ReconFormer: Accelerated MRI Reconstruction Using Recurrent Transformer

Pengfei Guo, Yiqun Mei, Jinyuan Zhou, Shanshan Jiang, Member, IEEE, and Vishal M. Patel, Senior Member, IEEE

Abstract—The accelerating magnetic resonance imaging (MRI) reconstruction process is a challenging ill-posed inverse problem due to the excessive under-sampling operation in k-space. In this paper, we propose a recurrent Transformer model, namely ReconFormer, for MRI reconstruction, which can iteratively reconstruct high-fidelity magnetic resonance images from highly under-sampled k-space data (e.g., up to 8x acceleration). In particular, the proposed architecture is built upon Recurrent Pyramid Transformer Layers (RPTLs). The core design of the proposed method is Recursive Scale-wise Attention (RSA), which jointly exploits intrinsic multi-scale information at every architecture unit as well as the dependencies of the deep feature correlation through recurrent states. Moreover, benefiting from its recurrent nature, ReconFormer is lightweight compared to other baselines and only contains 1.1 M trainable parameters. We validate the effectiveness of ReconFormer on multiple datasets with different magnetic resonance sequences and show that it achieves significant improvements over the state-of-the-art methods with better parameter efficiency. The implementation code and pre-trained weights are available at https://github.com/guopengf/ReconFormer.

Index Terms—MRI, reconstruction, deep learning, transformer.

I. INTRODUCTION

MAGNETIC Resonance Imaging (MRI) is one of the most prevalent diagnostic and research tools in clinical scenarios, which provides excellent resolution and abundant contrast mechanisms to visualize different structural and functional properties of the underlying anatomy. Due to physiological and hardware constraints [1], the MRI acquisition process is inherently slow. Consequently, extending acquisition time to collect complete data in k-space (frequency domain) imposes a significant burden on patients and makes MRI less accessible. One of the common approaches to accelerate the MRI acquisition procedure is to collect partial data by under-sampling in k-space. However, such an operation violates the Nyquist-Shannon sampling theorem and introduces aliasing artifacts in the reconstructed image. Compressed sensing (CS) mitigates this issue by formulating image reconstruction as an optimization problem with several assumptions, including sparsity and incoherence [2]. Past literature in advanced CS-based image reconstruction has exploited low-rank constraints [3], adaptive sparse modeling [4], [5], and parallel imaging [6]. CS algorithms have several deficiencies impeding their practicality in real-world applications. First, CS recovery algorithms require careful tuning of hyper-parameters, which is problem-specific and non-trivial. Second, due to the nature of iterative optimization, CS methods often demand longer running time to achieve desirable reconstruction quality. Third, the acceleration factors are generally below 3 for typical MR images and can be potentially increased to 8-12 with parallel imaging techniques [7], [8], when high-frequency oscillatory artifacts are not reduced adequately during the optimization process [9].

In contrast, recently advanced deep learning-based methods are gaining more attention for fast and accurate MRI reconstruction. Rather than explicitly defining prior information and regularization functions, networks learn them implicitly from data. Convolutional neural network (CNN)-based MRI reconstruction methods have been shown to provide much better MR image quality than conventional CS-based methods [10]. Such methods focus on novel architecture designs, including overcomplete networks [11], neural ODEs [12], invertible networks [13], cascaded architectures [14], [15], recurrent structures [10], [16], and diffusion models [17], [18], [19], [20], [21]. Although those methods achieve promising results, the basic convolution layer generally suffers from several drawbacks. First, a content-independent convolution kernel is not optimal for restoring different image regions [22], which is usually referred to as inductive bias (IB). CNNs incorporate an inductive bias that nearby pixels in the input image are more likely to be interrelated than those farther apart. This is commonly known as the locality bias and is used in designing convolutional filters with shared weights that are typically

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
small, enabling them to detect local patterns in the input image. On the other hand, Transformers lack strong inductive biases, making them more adaptable and less restrictive. This allows the model to explore a wider range of solutions and potentially find a more optimal solution [23]. Second, the long-range dependency is not effectively modeled in convolutional networks [24]. In this study, we present a Transformer-based MRI reconstruction model to overcome the strong inductive bias introduced by previous CNN-based MRI reconstruction approaches.

Recent advances in Transformer [24] models introduce the self-attention mechanism to capture global interactions between contexts and open a new possibility to tackle the challenges in MRI reconstruction. Although previous studies [29], [30], [31] demonstrated good performance, their architectures still heavily rely on the self-attention mechanism proposed in ViT [28] and cannot efficiently model the multi-scale information. A fundamental vision principle that visual elements vary at scales plays a vital role in computer vision [32] and MRI reconstruction [16]. To leverage this intrinsic property, previous vision Transformer models [25], [28], [30], [33], [34], [35] rely on the network topology, as shown in Fig. 1(a). However, such design suffers from the following. 1) Scale modeling is inefficient. Scale information is gradually processed through network hierarchy, so the multi-scale representation is only obtained at the end of stages. 2) Scale modeling is inflexible. Multi-scale design at the topology level is usually task-dependent and hard to generalize for other tasks.

In addition, achieving high performance with Transformer models typically necessitates many parameters, which may increase the difficulty of training the model effectively using limited MRI data. For example, MTrans [29] contains 673M parameters and consequently requires a large-scale dataset for training. Due to its recurrent nature, ReconFormer is relatively lightweight compared to other baseline models, having only 1.1 M trainable parameters. This makes it easier to train and deploy and reduces the risk of overfitting when working with smaller datasets. Additionally, the reduced complexity of ReconFormer means that it requires fewer computational resources, making it a more efficient option for practical applications.

In this paper, to overcome those issues, a novel Recurrent Transformer, termed ReconFormer, is proposed to recover the fully-sampled image from the under-sampled \( k \)-space data. In summary, the followings are our key contributions:

- ReconFormer introduces a powerful locally pyramidal but globally columnar architecture, which can perceive multi-scale representation at any stage while well preserving image details, as shown in Fig. 1(c).
- Recurrent Pyramid Transformer Layer (RPTL) is proposed to allow scale modeling at every basic building unit and exploit deep feature correlation through recurrent states.
- By fully employing the parameter efficiency of the recurrent structure, ReconFormer can be trained from scratch even with limited medical data. The effectiveness of our approach is validated on three datasets with various MR sequences and experiment results demonstrate that ReconFormer achieves prominent improvements over the state-of-the-art CNN-based as well as Transformer models with better parameter efficiency.

## II. RELATED WORKS

With recent advances in deep learning, deep network models are gaining more popularity for MR image reconstruction and show prominent performance in solving this challenging inverse problem. Several early approaches [36], [37], [38], proposed to use a single feed-forward CNN, such as SRCNN [39] and UNet [40], to learn an end-to-end mapping between the observed \( k \)-space and the fully-sampled data in the image domain. More advanced deep networks have exploited iterative architectures that decompose the end-to-end pipeline as several iterative learnable optimization stages. Sun et al. [41] proposed the ADMM-Net for MRI reconstruction, which formulates each deep network stage as an iteration of the alternating direction method of multipliers (ADMM) algorithm [42]. rsGAN [43] was proposed to conduct synergistic recovery of undersampled multi-contrast acquisitions using conditional generative adversarial networks. To imitate the iterative dictionary learning reconstruction methods, Schlemper et al. [14] and Dar et al. [44] proposed to leverage deep cascaded CNN architecture which performs the convolution operations in the image space by several cascaded residual learning blocks. Qin et al. [10] proposed a novel convolutional recurrent neural network (CRNN) architecture, which jointly utilizes the dependencies of the temporal sequences and the benefit of iterative optimization algorithms. Wang et al. [16], Chen et al. [12], and Guo et al. [11] further advanced the CRNN design by introducing the recurrent pyramid hierarchy, neural ODE, and overcomplete network architecture to achieve better reconstruction quality.

In light of recent advancements in the field of Transformers [28], several attempts [45], [46], [47] have been made for utilizing Transformers in addressing the intricate inverse problem associated with MR image reconstruction. Huang et al. [47] investigated the effectiveness of transformers for fast MRI reconstruction by comparing different novel network architectures and demonstrated that transformers perform well for MRI reconstruction under different undersampling conditions. The use of GAN’s adversarial structure improves the quality of reconstructed images. TRANS-Net [46] was proposed to emphasize the high-frequency information in MR images. Unlike traditional unrolled methods, TRANS-Net applies regularization not only on the reconstructed result but also on the error maps in the residual image domain. A transformer module is introduced to enhance the global spatial correlation of the reconstructed image for better tissue recovery. Huang et al. [45] propose a new Transformer architecture, SDAUT, for fast MRI that combines Shifted Windows Transformer with U-Net to reduce network complexity, as well as enhances the explainability of the reconstruction model by the deformable attention. Feng et al. [30] proposed a task Transformer network for...
Fig. 1. Comparisons of different Transformer architectures. (a) Recent high-level vision Transformer [25], [26] uses a pyramid structure to model multi-scale contexts. (b) Previous low-level vision Transformers [22], [27] and ViT [28] follow a simple columnar design, which preserves better information but fails to model at scales. (c) ReconFormer incorporates the pyramid structure inside a Transformer layer. The local pyramid enables scale processing at each unit while the globally columnar structure maintains high-resolution information. FFN refers to a feed-forward network in Transformer layers.

Fig. 2. The architecture of the proposed ReconFormer. (a) A schematic of unrolled ReconFormer iterations. Here, \( \otimes \) denotes the channel-wise concatenation, and RFB is the ReconFormer block. (b) The illustration of a recurrent unit (RU). Here, \( \oplus \) denotes the element-wise addition.

joint MRI reconstruction and super-resolution. MTrans [29] was proposed for multi-modal MRI reconstruction and super-resolution. Korkmaz et al. [35] proposed SLATER, which can perform self-supervised MRI reconstruction facilitated by an adversarial Transformer. A texture Transformer module was introduced in [31] for reference-based MRI reconstruction. Nonetheless, the issues of inefficient and inflexible scale modeling persist, as discussed in Section I. It is recognized that diffusion models have recently garnered significant interest within the medical image analysis community, exhibiting substantial promise for universal MRI reconstruction with quantified uncertainty. Chung et al. [21] presented a novel approach to accelerated MRI reconstruction using score-based diffusion models, which leverages the learned score function as the prior to sample data from a conditional distribution given the measurements. This model is trained with magnitude images only, yet it can reconstruct complex-valued data and even extend to parallel imaging. Quan et al. [20] proposed HGGDPRRec to leverage the denoising score matching and homotopic gradients of generative density priors (HGGDP) for accelerated MRI reconstruction. HGGDPRRec addresses the issues of low-dimensional manifold and low data density region in generative density prior by estimating the target gradients in a higher-dimensional space. The experiment results indicate that HGGDPRRec can generate high-quality images with highly under-sampled k-space data. Nevertheless, these models usually have extensive model sizes and substantial computational resource requirements for multiple samplings to achieve satisfactory reconstruction quality, which deviates from our primary objective of developing a lightweight model that is aptly suited for practical deployment.

Besides approaches focusing on MR reconstruction in the image domain, Han et al. [48] proposed a method \( k \)-space that focuses on restoring missing samples in the \( k \)-space domain. Later, KIKI-Net [15] was proposed to perform interleaved convolution operations on the image and \( k \)-space domains. Zhou et al. [49] proposed the DuDoRNet, demonstrating the potential of dual-domain learning for further improving reconstruction quality. While those previous CNN-based methods achieved prominent reconstruction performance, inherent drawbacks of convolution layers cannot be addressed well in a fully convolutional network design. Firstly, CNNs usually operate small convolution kernels, such as \( 3 \times 3 \), to scan the holistic image. Such interactions between images and convolution are content-independent, which promotes efficient training, but may not be optimal to restore different image regions [22]. In addition, due to the principle of local processing in CNNs, the long-range dependency is not effectively modeled, and the effective receptive field [50] size of CNNs is constrained under limited parameters [24]. Consequently, the leading methods in fastMRI leaderboard [51] require a large number of parameters to achieve good performance (e.g., i-RIM [52] (276M), VarNet [53] (30M), and PC-RNN [16] (23M)). In contrast,
the proposed ReconFormer only requires about 1M parameters by leveraging the parameter-efficient design and Transformer modules, similar to the weight-sharing strategy introduced in [54] to mitigate the computational burden.

While ViT [28] has achieved promising results on image classification without relying on convolutions. Its “columnar” design impedes modeling objects beyond the naive scale. To overcome this weakness, most recent Transformers (PVT [26], Swin Transformer [25], MViT [55], ViL [56]) for high-level vision adapt hierarchical structures, which can be view as a stack of multiple ViTs with reduced resolution. While they reveal considerable improvements over ViT[28], as discussed in section I, they still suffer from several limitations, and pyramid structures cannot be transferred for image restoration. Another way of scale processing is to be view as a stack of multiple ViTs with reduced resolution.

Another way of scale processing is to be view as a stack of multiple ViTs with reduced resolution.

As shown in Fig. 2(a), ReconFormer consists of three recurrent units and a refine module (RM). To maintain high-resolution information, ReconFormer employs a globally columnar structure. In particular, recurrent units map the input degraded images to the original dimension. Meanwhile, across each recurrent unit, the receptive fields of recurrent units are gradually reduced to reconstruct high-quality MR images coarse-to-fine. It is worth noting that the last recurrent unit RU employs complete architecture [11], which has been demonstrated to learn low-level features in MRI reconstruction effectively.

As shown in Fig. 2(b), a recurrent unit contains an encoder $f^{\text{Enc}}$, a ReconFormer block $f^{\text{RFB}}$, and a decoder $f^{\text{Dec}}$. The encoder and decoder are built upon convolution layers to have a more stable optimization [58] and better early visual processing [22]. We denote a single recurrent unit RU, where $f_i$ where $i \in \{1, 2, 3\}$ indicates the $i^{\text{th}}$ recurrent unit and its operation in the $i^{\text{th}}$ iteration can be formulated as follows:

$$f_{i}^{\text{RU}}(y_{i-1}^{(t)}, h_{i}^{(t)}, c_{i}^{(t)}) = f^{\text{Dec}}(f_{i}^{\text{RFB}}(h_{i}^{(t)}, c_{i}^{(t)}) \oplus f_{i}^{\text{Enc}}(y_{i-1}^{(t)})),$$

(4)

where $\oplus$ denotes the element-wise addition. ReconFormer blocks $f^{\text{RFB}}$ in three RUs operate on features at different scale sizes (i.e., $\times 0.5, \times 1.0, \times 1.5$, respectively) of the original image by adjusting the stride convolution layers in encoders $f^{\text{Enc}}$. Then, the decoder $f^{\text{Dec}}$ in each RU uses transposed convolution layers to recover the original image size. A data consistency (DC) layer is added at the end of each decoder network to reinforce the data consistency in the $k$-space as follows:

$$\hat{y}_{i}^{(t)} = \text{DC}(f_{i}^{\text{RU}}(y_{i-1}^{(t)}, h_{i}^{(t)}, c_{i}^{(t)}), x, U),$$

$$= F^{-1}[UX + (1 - U)F(\hat{y}_{i}^{(t)} R_{F}^{i})],$$

(5)

where $U$ is a binary under-sampling mask in $F_{u}$ which is used for data consistency in $k$-space. $(\hat{y}_{i}^{(t)}, \hat{y}_{i}^{(t)}) \in \mathbb{R}^{H \times W \times 2}$ are input and output of RU, respectively, where $H$ and $W$ are the zero-filled reconstruction height and width, and 2 represents the real and imaginary channels. $h_{i}^{(t)} \in \mathbb{R}^{\lambda \times \lambda \times C}$ is the hidden state from the previous iteration, where $\lambda$ is the scaling factor determined by the structure of each RU and consequently controls the size of receptive fields. $c_{i}^{(t)}$ is the deep feature correlation from the previous iteration. At the $i^{\text{th}}$ iteration, ReconFormer consists of three recurrent units and can be unrolled as follows:

$$\hat{y}_{1}^{(t)} = \text{RU}_{1}(\hat{y}_{0}^{(t)}, h_{1}^{(t)}, c_{1}^{(t)}, x, U),$$

$$\hat{y}_{2}^{(t)} = \text{RU}_{2}(\hat{y}_{1}^{(t)}, h_{2}^{(t)}, c_{2}^{(t)}, x, U),$$

$$\hat{y}_{3}^{(t)} = \text{RU}_{3}(\hat{y}_{2}^{(t)}, h_{3}^{(t)}, c_{3}^{(t)}, x, U),$$

$$\hat{y}_{t+1}^{(t)} = \text{DC}(\text{RM}(\hat{y}_{1}^{(t)} \otimes \hat{y}_{2}^{(t)} \otimes \hat{y}_{3}^{(t)}), x, U),$$

(6)

where $\otimes$ denotes the channel-wise concatenation. Leveraging coarse-to-fine reconstructions in different receptive fields can further improve image fidelity. Similar to [11], [16], the RM fuses the outputs of recurrent units to generate $\hat{y}_{t+1}^{(t)}$, which is the output of the current iteration and also servers as the input for the next iteration.

B. ReconFormer

As shown in Fig. 2(a), ReconFormer consists of three recurrent units and a refine module (RM). To maintain high-resolution information, ReconFormer employs a globally columnar structure. In particular, recurrent units map the input degraded images to the original dimension. Meanwhile, across each recurrent unit, the receptive fields of recurrent units are gradually reduced to reconstruct high-quality MR images coarse-to-fine. It is worth noting that the last recurrent unit RU employs complete architecture [11], which has been demonstrated to learn low-level features in MRI reconstruction effectively.

As shown in Fig. 2(b), a recurrent unit contains an encoder $f^{\text{Enc}}$, a ReconFormer block $f^{\text{RFB}}$, and a decoder $f^{\text{Dec}}$. The encoder and decoder are built upon convolution layers to have a more stable optimization [58] and better early visual processing [22]. We denote a single recurrent unit RU, where $f_{i}$ where $i \in \{1, 2, 3\}$ indicates the $i^{\text{th}}$ recurrent unit and its operation in the $i^{\text{th}}$ iteration can be formulated as follows:

$$f_{i}^{\text{RU}}(y_{i-1}^{(t)}, h_{i}^{(t)}, c_{i}^{(t)}) = f^{\text{Dec}}(f_{i}^{\text{RFB}}(h_{i}^{(t)}, c_{i}^{(t)}) \oplus f_{i}^{\text{Enc}}(y_{i-1}^{(t)}),$$

(4)

where $\oplus$ denotes the element-wise addition. ReconFormer blocks $f^{\text{RFB}}$ in three RUs operate on features at different scale sizes (i.e., $\times 0.5, \times 1.0, \times 1.5$, respectively) of the original image by adjusting the stride convolution layers in encoders $f^{\text{Enc}}$. Then, the decoder $f^{\text{Dec}}$ in each RU uses transposed convolution layers to recover the original image size. A data consistency (DC) layer is added at the end of each decoder network to reinforce the data consistency in the $k$-space as follows:

$$\hat{y}_{i}^{(t)} = \text{DC}(f_{i}^{\text{RU}}(y_{i-1}^{(t)}, h_{i}^{(t)}, c_{i}^{(t)}), x, U),$$

$$= F^{-1}[UX + (1 - U)F(\hat{y}_{i}^{(t)} R_{F}^{i})],$$

(5)

where $U$ is a binary under-sampling mask in $F_{u}$ which is used for data consistency in $k$-space. $(\hat{y}_{i}^{(t)}, \hat{y}_{i}^{(t)}) \in \mathbb{R}^{H \times W \times 2}$ are input and output of RU, respectively, where $H$ and $W$ are the zero-filled reconstruction height and width, and 2 represents the real and imaginary channels. $h_{i}^{(t)} \in \mathbb{R}^{\lambda \times \lambda \times C}$ is the hidden state from the previous iteration, where $\lambda$ is the scaling factor determined by the structure of each RU and consequently controls the size of receptive fields. $c_{i}^{(t)}$ is the deep feature correlation from the previous iteration. At the $i^{\text{th}}$ iteration, ReconFormer consists of three recurrent units and can be unrolled as follows:

$$\hat{y}_{1}^{(t)} = \text{RU}_{1}(\hat{y}_{0}^{(t)}, h_{1}^{(t)}, c_{1}^{(t)}, x, U),$$

$$\hat{y}_{2}^{(t)} = \text{RU}_{2}(\hat{y}_{1}^{(t)}, h_{2}^{(t)}, c_{2}^{(t)}, x, U),$$

$$\hat{y}_{3}^{(t)} = \text{RU}_{3}(\hat{y}_{2}^{(t)}, h_{3}^{(t)}, c_{3}^{(t)}, x, U),$$

$$\hat{y}_{t+1}^{(t)} = \text{DC}(\text{RM}(\hat{y}_{1}^{(t)} \otimes \hat{y}_{2}^{(t)} \otimes \hat{y}_{3}^{(t)}), x, U),$$

(6)

where $\otimes$ denotes the channel-wise concatenation. Leveraging coarse-to-fine reconstructions in different receptive fields can further improve image fidelity. Similar to [11], [16], the RM fuses the outputs of recurrent units to generate $\hat{y}_{t+1}^{(t)}$, which is the output of the current iteration and also servers as the input for the next iteration.
Fig. 3. (a) The schematic of recurrent scale-wise attention (RSA). (b) An illustration of the transition function in the RPTL. Here, $\otimes$ and $\oplus$ denote the element-wise multiplication and addition. (c) The architecture of the proposed RPTL.

Fig. 4. A schematic of unrolled ReconFormer block (RFB). Here, we show the RFB only consisting of a single recurrent pyramid Transformer layer for simplicity.

C. Recurrent Pyramid Transformer Layer

A ReconFormer block (RFB) is formed by stacked Recurrent Pyramid Transformer Layers (RPTLs). The core design of RPTL is the Recurrent Scale-wise Attention (RSA), as shown in Fig. 3(a). Compared to the standard multi-head self-attention of the original Transformer layer [28], the main difference lies in two aspects: (i) instead of performing a single attention function, the proposed RSA consists of several attention scale heads (i.e., $\times 1$, $\times 3$, and $\times 5$), which operate on multi-scale patches in parallel. Such design enables efficient in-place scale modeling and forms a feature pyramid by directly projecting features at various scales into multiple attention heads. Consequently, the proposed RPTL allows scale processing at basic architecture units. (ii) The correlation estimation in the proposed RSA relies on both the hidden state $h(t)$ and the deep feature correlation $c(t)$ from the previous iteration, which enables more robust correlation estimation by propagating correlation information between adjacent states.

The proposed RPTL utilizes a local attention mechanism with a shifted window scheme [25], which brings greater efficiency and shows superiority on several tasks [22], [59], [60]. Given the input hidden state $h^{(t)} \in \mathbb{R}^{H \times W \times C'}$, RPTL first uses non-overlapping $M \times M$ local windows to partition $h^{(t)}$ into a feature of size $H' \times W' \times M^2 \times C'$, where $M$ is the size of local windows and $\frac{H' \times W'}{M^2}$ is the number of windows. Then, the proposed RSA operates on each window in parallel to compute the self-attention. Let $I \in \mathbb{R}^{M^2 \times C'}$ denote a local window feature, the query $Q$, key $K$, and value $V$ matrices in RPTL are formulated as follows:

$$Q = I P_Q, K = I P_K, V = I P_V,$$

where $P_Q$, $P_K$, and $P_V$ are the corresponding projection matrices in attention scale heads. As shown in Fig. 3(b), we compute the outputs of RSA as follows:

$$\text{Attention}(Q, K, V, c^{(t)}) = \text{SoftMax}(c^{(t+1)}) V.$$  

At each time step $t$, the deep feature correlation $c^{(t)}$ is updated recursively as follows:

$$c^{(t+1)} = \lambda \left( \frac{Q K^T}{\sqrt{d}} + B \right) + (1 - \lambda)c^{(t)},$$

where $d$ is the embedding dimension. $B$ is a learnable relative position bias, and $\lambda$ is a learnable weighting factor controlling the contribution of the deep feature correlation from the adjacent state. As shown in Fig. 3(c), we can formulate the workflow of RPTL as follows:

$$I = \text{RSA} (\text{LN}(I)) + I,$$

$$I = \text{MLP} (\text{LN}(I)) + I,$$

where LN and MLP denote the layer normalization and the multi-layer perceptron, respectively. The schematic of a ReconFormer block is shown in Fig. 4.

IV. EXPERIMENTS AND RESULTS

A. Datasets and Implementation Details

The SKM-TEA [64], fastMRI [51], and HPKS [65] datasets are used for conducting experiments. The SKM-TEA [64] raw data track provides 155 complex-valued...
TABLE I

| Method | Param. | FLOPS | NSMI | SSIM | PSNR | NMSE | GMSE |
|--------|--------|-------|------|------|------|------|------|
| CR [61] | - | - | 0.0399 ± 0.002 | 0.3040 ± 0.002 | 0.7155 ± 0.002 | 0.7155 ± 0.002 | 29.41 ± 1.74 |
| Net [10] | 1.855M | 332 | 0.0300 ± 0.000 | 0.0434 ± 0.003 | 0.1255 ± 0.000 | 0.3268 ± 0.007 | 34.01 ± 1.97 |
| RAMNet [15] | 2.060M | 332 | 0.0416 ± 0.004 | 0.0463 ± 0.003 | 0.1163 ± 0.000 | 0.4614 ± 0.008 | 35.00 ± 2.50 |
| ReconFormer | 1.515M | 4440 | 0.0397 ± 0.002 | 0.0368 ± 0.003 | 0.1276 ± 0.000 | 0.4996 ± 0.009 | 34.02 ± 1.99 |
| DSC5 [14] | 2.274M | 332 | 0.0400 ± 0.003 | 0.0365 ± 0.002 | 0.1555 ± 0.003 | 0.4629 ± 0.008 | 37.91 ± 2.57 |
| OCR [11] | 1.193M | 332 | 0.0407 ± 0.003 | 0.0234 ± 0.003 | 0.1747 ± 0.008 | 0.3964 ± 0.034 | 39.33 ± 2.50 |
| ReconFormer | 1.515M | 5280 | 0.0500 ± 0.003 | 0.0375 ± 0.003 | 0.1574 ± 0.003 | 0.5103 ± 0.034 | 40.00 ± 2.55 |

Fig. 5. Effective receptive fields [50] (ERFs) of different CNN-based methods and the proposed ReconFormer. ERFs are computed at the last layer of the model w.r.t. the center. For a better visualization in the peripheral region with lower gradient magnitudes, we apply $\gamma$ correction with $\gamma = 0.1$ for all generated ERFs [61].

multi-coil $T_2$-weighted (qDESS [66]) knee MRI scans. 124, 10, and 21 coil-combined volumes are used for training, validation, and testing. Each subject provides approximately 160 cross-sectional knee images with the matrix of size 512 × 512. The fastMRI [51] dataset contains 1,172 complex-valued single-coil coronal proton density (PD)-weighted knee MRI scans. Each scan provides approximately 35 coronal cross-sectional knee images with the matrix of size 320 × 320. We partition this dataset into 973 scans for training and 199 scans (fastMRI validation dataset) for testing. The HPKS [65] dataset provides complex-valued single-coil axial $T_1$-weighted brain MRI scans from 144 post-treatment patients with malignant glioma. Each scan contains 15 axial cross-sectional images with a matrix of size 256 × 256. The data splits are as follows: 102 scans are used for training, 14 scans are used for validation, and 28 scans are used for testing. In comparison experiments, the input under-sampled image sequences are generated by randomly under-sampling the k-space data using the 1D uniform-density under-sampling function similar to the fastMRI challenge [51].

Models are trained using the $\ell_1$ loss with Adam optimizer with the following hyperparameters: learning rate of $2.0 \times 10^{-4}$; 50 maximum epochs; the batch size of 4; the number of recurrent iterations $T$ of 5; the size of local windows $M$ of 8. Peak signal-to-noise ratio (PSNR), structural index similarity (SSIM), and normalized mean square error (NMSE) are used as evaluation metrics for comparison. Performance comparison tables report metrics’ mean and standard deviation across test subjects. We evaluate the statistical significance of performance differences using Wilcoxon sign-rank tests. We implement the proposed model using PyTorch on NVIDIA RTX8000 GPUs. For a fair and comprehensive comparison with image restoration literature, we also include non-MR reconstruction methods into baselines. Those methods are modified for data with real and imaginary channels, and a DC layer is added at the end of the networks. The unrolling length of all iterative approaches is set to 5.

B. Comparisons With the State-of-the-Art

We compare the proposed method with seven representative methods, including conventional CS-based method [62], popular CNN-based methods – UNet [40], KIKI-Net [15], Kiu-net [63], and D5C5 [14], state-of-the-art iterative reconstruction approaches – OCR [11], and vision Transformer model – SwinIR [22].

1) Effective Receptive Fields: In Fig. 5, we plot effective receptive fields [50] (ERFs) of four representative methods, which represent how input pixels affect the center pixels at the last layer of the model. ERFs can be visualized by injecting the gradient to a unit in a particular layer and back-propagating it to the input [67], [68]. As shown in Fig. 5, ERFs of different methods are generally Gaussian-like. While the theoretical receptive fields of most methods can span the whole image, the pure CNN-based methods have a limited ERF. We can observe that gradient magnitudes rapidly decay from the center in ERFs of UNet, D5C5, and OCR. This means that only limited input pixels can effectively reach the output. However, we can observe a more widely distributed darker area in the ERF of the proposed ReconFormer, which indicates that Transformer-based models have relatively larger ERFs that span over the entire image. This validates one of our motivations that long-range dependency can be efficiently modeled in Transformers.
2) Quantitative Evaluations: Table I shows the quantitative results evaluated on the two datasets for AF=4 and AF=8. Compared with the other methods, the proposed ReconFormer achieves competitive performance on two datasets for all acceleration factors while containing the least number of parameters. It is worth noting that our approach exhibits a larger performance improvement when the acceleration factor increases (i.e., more challenging scenarios). In particular,
Fig. 7. Qualitative comparison of different methods on the multi-coil SKM-TEA dataset. The second column of each subplot shows the corresponding ×8 zoomed-in image. The red arrows point to where ReconFormer achieves high visual acuity.

**TABLE II**

| Method            | SKM-TEA |       |       |       |       |
|-------------------|---------|-------|-------|-------|-------|
|                   | NMSE    | AF = 4 | AF = 8 | AF = 4 | AF = 8 |
| UNet [40]         | 0.0204±0.003 | 0.0270±0.004 | 0.8469±0.026 | 0.7904±0.030 | 33.91±1.39 | 31.44±1.24 |
| KIKI-Net [15]     | 0.0196±0.003 | 0.0271±0.004 | 0.8577±0.026 | 0.7941±0.032 | 34.26±1.42 | 31.42±1.28 |
| Kiu-net [63]      | 0.0195±0.003 | 0.0264±0.004 | 0.8582±0.025 | 0.7942±0.032 | 34.31±1.40 | 31.65±1.27 |
| SwinIR [22]       | 0.0192±0.003 | 0.0256±0.003 | 0.8597±0.025 | 0.8022±0.031 | 34.45±1.37 | 31.94±1.27 |
| DSC5 [14]         | 0.0188±0.003 | 0.0257±0.004 | 0.8648±0.026 | 0.8030±0.032 | 34.63±1.44 | 31.89±1.30 |
| OUCR [11]         | 0.0183±0.003 | 0.0246±0.003 | 0.8693±0.025 | 0.8113±0.032 | 34.89±1.45 | 32.27±1.32 |
| ReconFormer       | 0.0179±0.003 | 0.0239±0.003 | **0.8730±0.025** | **0.8158±0.031** | **35.06±1.45** | **32.51±1.32** |

Fig. 8. k-space analysis of (a) low frequency and (b) high frequency on the HPKS dataset. (c) Unrolling length v.s. Reconstruction performance of ReconFormer on HPKS dataset with 4× acceleration.

for the HPKS and fastMRI datasets, our model outperforms the most competitive approach OUCR [11] by 0.90 dB and 0.30 dB in 8× acceleration compared to 0.76 dB and 0.12 dB in 4× acceleration, respectively. Table II
shows the quantitative results evaluated on the multi-coil SKM-TEA dataset. We can observe that the superiority of ReconFormer remains across different tissue contrasts, acceleration rates, and acquisition settings. While the absolute performance improvement is distinct in each dataset is various due to different acquisition quality, all reported improvements achieved by ReconFormer are statistically significant ($p < 0.05$).

3) Qualitative Evaluations: Fig. 6 shows qualitative evaluations of two cases on HPKS and fastMRI. It can be seen that ReconFormer yields better reconstruction quality and perceptually shows improved artifact suppression compared to other methods, as shown in the red boxes of each sub-figure. In Fig. 7, we present two cases from the multi-coil SKM-TEA dataset and their corresponding 8×8 zoomed-in images. The proposed ReconFormer can provide reconstruction with more apparent sharper tissue contrast and high visual over. We conduct the $k$-space analysis to investigate the performance of different methods in low (1/3 center frequencies) and high (2/3 peripheral frequencies) spatial frequencies. The center of the $k$-space corresponds to low-frequency information (general shape and contrast), while the peripheral region corresponds to high-frequency information (details and edges). By conducting a $k$-space analysis, we can quantitatively assess how well different methods can preserve and reconstruct the information from under-sampled $k$-space data. This is particularly important in MRI reconstruction, where low and high-frequency information is crucial for accurate diagnosis. By comparing the reconstructed $k$-space of different methods, we can gain insights into their performance that may not be apparent from the reconstructed images alone. As shown in Fig. 8, the proposed method is shown to be more effective for both low-frequency information and high-frequency features. The superior performance on different MRI sequences demonstrates the merit of the proposed Transformer architecture and our RPTL in jointly utilizing intrinsic multi-scale information and the deep feature correlation.

### C. Ablation Study

1) Architecture Design: We first separately evaluate the effectiveness of network components, including three recurrent units, RM and PRTL, in the proposed ReconFormer. As shown in Table III, when we add each recurrent unit, the reconstruction quality can be gradually boosted, demonstrating the architecture design’s effectiveness. Then, we show that adding RM to fuse the coarse-to-fine reconstructions further improves the performance. Moreover, RPTL consistently provides performance improvements in different acceleration factors, which implies the importance of fully exploiting multi-scale information in basic building blocks. For experiments without RPTL, we replace RSA with the standard shifted window self-attention [25].

As discussed in Sec. III-C, the major difference between the standard multi-head self-attention and the proposed RSA is that our attention module enables efficient in-place scale modeling and forms a feature pyramid by multiple scale attention heads. To investigate the influences of pyramid scales, we fix all other hyperparameters and further conduct an ablation study by gradually using more scales in the formed feature pyramid. As shown in Table IV, performance gains can be constantly obtained when more pyramid scales are added. Since attention searching space in the feature pyramid is gradually expanded to more scales, there is a better chance for the attention unit to find more informative correspondences. These results validate one of our motivations that modeling multi-scale relationships is critical in image restoration. While the best performance is achieved when all four scales are included, the model size and FLOPs are also increased from 0.94M and 301G to 1.25M and 360G, respectively, compared to only using ×1 attention scale head. Hence, three attention scale heads (i.e., ×1, ×3, and ×5) are used in the proposed RSA to balance performance and efficiency. In addition, we investigate the influence of the size of local windows and $M$ and the number of recurrent iterations $T$ on reconstruction results in Fig. 9. The relationship between reconstruction accuracy and computational efficiency is intertwined, and the vision transformers employ a shifted windowing scheme to balance the trade-off. This scheme restricts the self-attention computation to non-overlapping local windows, enabling greater efficiency while also permitting a cross-window connection. Although enlarging the window size may enhance performance, it also increases resource usage considerably. As depicted in Fig. 9(a), expanding the window size to 16 ($M=16$) leads to the memory usage of approximately 35GB for processing a single data point during training. In Fig. 9(b), We can observe that reconstruction quality increases as the number of recurrent iterations expands but gets saturated when $T = 5$.

2) Generalizability: We note that DL-based methods usually train dedicated models using fixed acceleration factors and under-sampling patterns. To examine the robustness of dedicated models on various under-sampling operations, we conduct control experiments in Table V, where the under-sampled Fourier encoding matrix $F_u$ is mismatched between training and inference. We first investigate the robustness of different models on different acceleration factors, which implies the importance of fully exploiting multi-scale information in basic building blocks. For experiments without RPTL, we replace RSA with the standard shifted window self-attention [25].

| Modules | AF=4 | AF=8 |
|---------|------|------|
| RU1     | 38.19| 31.39|
| RU2     | 39.63| 32.47|
| RU3     | 39.79| 32.75|
| RM      | 39.94| 32.90|
| RPTL    | 40.09| 33.04|

| Attention Scale Heads | Param. | FLOPs | AF=4 | AF=8 |
|-----------------------|--------|-------|------|------|
| ×1                    | 0.94M  | 301G  | 39.80| 32.74|
| ×3                    | 1.09M  | 332G  | 39.92| 32.88|
| ×5                    | 1.14M  | 342G  | 40.09| 33.04|
| ×7                    | 1.25M  | 360G  | 40.17| 33.15|

**Table III:** The Ablation Study of Proposed Modules on HPKS

**Table IV:** The Ablation Study of Attention Scale Heads on HPKS.
The potential of using Transformer-based models to recover high-fidelity MR images from under-sample data is investigated in this study. Our findings indicate that our proposed ReconFormer, which incorporates Recurrent Scale-wise Attention, can effectively enhance reconstruction quality and efficiently model multi-scale information in each layer. Based on our experimental results, our proposed approach offers several advantages. First, the architecture design of ReconFormer utilizes a locally pyramidal yet globally columnar structure, modeling for multi-scale representation at any stage in the process while maintaining image details with great accuracy. This unique architecture is particularly effective for image restoration tasks, such as MRI reconstruction demonstrated in this paper. Second, a new self-attention mechanism, Recurrent Scale-wise Attention, is introduced in the proposed RPTL to achieve scale modeling at each fundamental building block and leverage the deep feature correlation via recurrent states. The RPTL utilizes a pyramidal structure that enables it to perform multi-scale modeling of input sequences at every level of abstraction. This allows for more precise modeling of complex input sequences that may contain information at varying levels of granularity. The RPTL also incorporates a recurrent state mechanism that will enable it to propagate the correlation between adjacent states, which helps to exploit the temporal dynamics of sequential data and can lead to improved model performance. Third, ReconFormer, with only 1.1 M trainable parameters due to its recurrent nature, is relatively lightweight compared to other baseline models. For instance, MTrans [29] has 673M parameters, necessitating a large-scale dataset for practical training. This feature makes ReconFormer more manageable to train and deploy and minimizes the risk of overfitting when working with smaller datasets in practical applications.

Although our proposed method demonstrated competitive performance, there are a few limitations that need to be taken into consideration when interpreting the results of our study. First, we notice that the absolute performance improvement achieved by ReconFormer varies across different datasets. The difference in performance between datasets is mainly due to variations in the quality of the images acquired, resulting in differing levels of difficulty in reconstructing the images. It is worth noting that all baseline methods suffer into consideration when interpreting the results of our study. First, we notice that the absolute performance improvement achieved by ReconFormer varies across different datasets. The difference in performance between datasets is mainly due to variations in the quality of the images acquired, resulting in differing levels of difficulty in reconstructing the images. It is worth noting that all baseline methods suffer from this common issue. For example, although OUCR [11] is the most competitive method, it can only achieve a 0.38 dB improvement in SKM-TEA, while it outperforms D5C5 [14] by 1.74 dB in HPKS (8× acceleration). To address this issue, some studies, such as Peng et al. [18], remove slices with little information (e.g., the first five or ten slices) from the evaluation for fastMRI and SKM-TEA.

Second, the presence of out-of-distribution input data can potentially hinder the performance of ReconFormer, as shown in the last column of Table V, we apply models trained on Cartesian sampling on Varden-sampled input when the acceleration factor is set to 4. We observe that the cascaded model (D5C5) and iterative methods (OUCR and ReconFormer) are more robust to such sampling pattern mismatch. This is mainly because those methods carry out more data consistency in each cascaded block or iterative step, which forces deep networks to learn to recover missing samples in k-space. These results indicate that the generalizability of ReconFormer is prominent, and performance gains brought by the proposed method are consistent under various out-of-distribution under-sampling operations.

### V. Discussion

The potential of using Transformer-based models to recover high-fidelity MR images from under-sample data is investigated in this study. Our findings indicate that our proposed ReconFormer, which incorporates Recurrent Scale-wise Attention, can effectively enhance reconstruction quality and efficiently model multi-scale information in each layer. Based on our experimental results, our proposed approach offers several advantages. First, the architecture design of ReconFormer utilizes a locally pyramidal yet globally columnar structure, modeling for multi-scale representation at any stage in the process while maintaining image details with great accuracy. This unique architecture is particularly effective for image restoration tasks, such as MRI reconstruction demonstrated in this paper. Second, a new self-attention mechanism, Recurrent Scale-wise Attention, is introduced in the proposed RPTL to achieve scale modeling at each fundamental building block and leverage the deep feature correlation via recurrent states. The RPTL utilizes a pyramidal structure that enables it to perform multi-scale modeling of input sequences at every level of abstraction. This allows for more precise modeling of complex input sequences that may contain information at varying levels of granularity. The RPTL also incorporates a recurrent state mechanism that will enable it to propagate the correlation between adjacent states, which helps to exploit the temporal dynamics of sequential data and can lead to improved model performance. Third, ReconFormer, with only 1.1 M trainable parameters due to its recurrent nature, is relatively lightweight compared to other baseline models. For instance, MTrans [29] has 673M parameters, necessitating a large-scale dataset for practical training. This feature makes ReconFormer more manageable to train and deploy and minimizes the risk of overfitting when working with smaller datasets in practical applications.

Although our proposed method demonstrated competitive performance, there are a few limitations that need to be taken into consideration when interpreting the results of our study. First, we notice that the absolute performance improvement achieved by ReconFormer varies across different datasets. The difference in performance between datasets is mainly due to variations in the quality of the images acquired, resulting in differing levels of difficulty in reconstructing the images. It is worth noting that all baseline methods suffer from this common issue. For example, although OUCR [11] is the most competitive method, it can only achieve a 0.38 dB improvement in SKM-TEA, while it outperforms D5C5 [14] by 1.74 dB in HPKS (8× acceleration). To address this issue, some studies, such as Peng et al. [18], remove slices with little information (e.g., the first five or ten slices) from the evaluation for fastMRI and SKM-TEA.

Second, the presence of out-of-distribution input data can potentially hinder the performance of ReconFormer, as shown in the last column of Table V, we apply models trained on Cartesian sampling on Varden-sampled input when the acceleration factor is set to 4. We observe that the cascaded model (D5C5) and iterative methods (OUCR and ReconFormer) are more robust to such sampling pattern mismatch. This is mainly because those methods carry out more data consistency in each cascaded block or iterative step, which forces deep networks to learn to recover missing samples in k-space. These results indicate that the generalizability of ReconFormer is prominent, and performance gains brought by the proposed method are consistent under various out-of-distribution under-sampling operations.

### V. Discussion

The potential of using Transformer-based models to recover high-fidelity MR images from under-sample data is investigated in this study. Our findings indicate that our proposed ReconFormer, which incorporates Recurrent Scale-wise Attention, can effectively enhance reconstruction quality and efficiently model multi-scale information in each layer. Based on our experimental results, our proposed approach offers several advantages. First, the architecture design of ReconFormer utilizes a locally pyramidal yet globally columnar structure, modeling for multi-scale representation at any stage in the process while maintaining image details with great accuracy. This unique architecture is particularly effective for image restoration tasks, such as MRI reconstruction demonstrated in this paper. Second, a new self-attention mechanism, Recurrent Scale-wise Attention, is introduced in the proposed RPTL to achieve scale modeling at each fundamental building block and leverage the deep feature correlation via recurrent states. The RPTL utilizes a pyramidal structure that enables it to perform multi-scale modeling of input sequences at every level of abstraction. This allows for more precise modeling of complex input sequences that may contain information at varying levels of granularity. The RPTL also incorporates a recurrent state mechanism that will enable it to propagate the correlation between adjacent states, which helps to exploit the temporal dynamics of sequential data and can lead to improved model performance. Third, ReconFormer, with only 1.1 M trainable parameters due to its recurrent nature, is relatively lightweight compared to other baseline models. For instance, MTrans [29] has 673M parameters, necessitating a large-scale dataset for practical training. This feature makes ReconFormer more manageable to train and deploy and minimizes the risk of overfitting when working with smaller datasets in practical applications.

Although our proposed method demonstrated competitive performance, there are a few limitations that need to be taken into consideration when interpreting the results of our study. First, we notice that the absolute performance improvement achieved by ReconFormer varies across different datasets. The difference in performance between datasets is mainly due to variations in the quality of the images acquired, resulting in differing levels of difficulty in reconstructing the images. It is worth noting that all baseline methods suffer from this common issue. For example, although OUCR [11] is the most competitive method, it can only achieve a 0.38 dB improvement in SKM-TEA, while it outperforms D5C5 [14] by 1.74 dB in HPKS (8× acceleration). To address this issue, some studies, such as Peng et al. [18], remove slices with little information (e.g., the first five or ten slices) from the evaluation for fastMRI and SKM-TEA.

Second, the presence of out-of-distribution input data can potentially hinder the performance of ReconFormer, as shown
in Table V. Generally, deep learning methods require the development of a dedicated model tailored to a particular MRI sequence and under-sampling pattern. For example, if a model designed for four times acceleration is used for eight times acceleration input, its performance may considerably decline. Developing a universal model that can handle various MRI sequences and under-sampling patterns is an intriguing area of research. We are aware that diffusion models have recently gained considerable attention in the medical image analysis community and have demonstrated outstanding potential for universal MRI reconstruction [19], [20], [21]. We acknowledge that diffusion models have shown promising MRI reconstruction performance in different settings [17], [18]. Nevertheless, one of the primary goals of this study is to create a lightweight model suitable for efficient deployment. Given this focus, these models, while effective in their own right, have not been included in our comparative experiments due to their extensive model size and substantial computational resource requirements for multiple samplings to achieve satisfactory reconstruction quality, which diverges from our targeted scenarios. Third, the performance of the proposed method can be further improved using a larger size of local windows and more recurrent iteration when resource usage is not a constraint. In Fig. 9, we examine the impact of the number of recurrent iterations \( T \) and the local window size \( M \) on the reconstruction quality. Our results demonstrate that as \( T \) increases, the reconstruction quality improves; however, it reaches a saturation point when \( T = 5 \). Balancing reconstruction accuracy and computational efficiency remains a challenge. The shifted windowing scheme in vision transformers aims to optimize efficiency by restricting self-attention computation to non-overlapping local windows while maintaining a cross-window connection. Although increasing the window size can enhance performance, it can also impose a significant resource usage burden. As shown in Fig. 9(a), when the window size \( M \) is increased to 16, the memory usage required to process a single data point during training is approximately 35GB. These results highlight the importance of carefully selecting hyperparameters, such as \( T \) and \( M \), to balance reconstruction accuracy and computational efficiency, especially in resource-constrained environments.

VI. CONCLUSION

In this paper, we propose a recurrent Transformer-based MRI reconstruction model ReconFormer. By leveraging the novel RPTL, we can explore the multi-scale representation at every basic building unit and discover the dependencies of the deep feature correlation between adjacent recurrent modules. Incorporating three recurrent units and the refine module, ReconFormer reconstructs high-quality MR images through a locally pyramidal but globally columnar structure and achieves state-of-the-art performance on multiple datasets. ReconFormer is lightweight and does not require pre-training on large-scale datasets. Our experiments suggest the promising potential of using Transformer-based models in the MRI reconstruction task.
