Abstract. This paper considers the problem of fast MRI reconstruction. We propose a novel Transformer-based framework for directly processing the sparsely sampled signals in k-space, going beyond the limitation of regular grids as ConvNets do. We adopt an implicit representation of spectrogram, treating spatial coordinates as inputs, and dynamically query the partially observed measurements to complete the spectrogram, i.e. learning the inductive bias in k-space. To strive a balance between computational cost and reconstruction quality, we build an hierarchical structure with low-resolution and high-resolution decoders respectively. To validate the necessity of our proposed modules, we have conducted extensive experiments on two public datasets, and demonstrate superior or comparable performance over state-of-the-art approaches. Project page: https://zhaoziheng.github.io/Website/K-Space-Transformer

Keywords: MRI reconstruction · K-space completion · Transformer

1 Introduction

Magnetic Resonance Imaging (MRI) has been widely adopted as an efficient and non-invasive approach for routine examination and diagnosis. In general, the signals are collected, digitized and plugged into k-space for image reconstruction, i.e. an array of numbers representing spatial frequencies in the MR image. However, one challenge is, the full signal acquisition can be time-consuming, limiting its use on high resolution or dynamic imaging. Accordingly, various techniques have been developed for MR image reconstruction under the condition of incomplete observations.

For the last two decades, compressed sensing (CS) has been the major breakthrough in this field, allowing to reconstruct k-space signals from only partial measurements, however, the iterative optimization process often suffers from heavy parameter tuning [17] and long processing time [6]. Till recently, deep learning has been widely adopted for its better reconstruction performance and real-time imaging [18]. Generally speaking, the existing works can be cast into two paradigms, one focuses on reconstruction in image domain [7,12,15,4], with the k-space information being used in data consistency layer or loss function. The
other line of research takes k-space reconstruction into consideration, for example, integrate it to the normal image reconstruction pipeline in a recurrent or parallel manner [3,13,19,10]; or perform reconstruction purely in k-space [5,2].

In existing approaches on k-space reconstruction, almost all the architectures are based on convolutional neural networks (CNNs), with two inevitable limitations: First, the inductive bias of CNNs is fundamentally not suitable for spectrogram, for example, the kernels in CNNs are normally shared across all spatial positions, i.e., equivariance assumption, however, on k-space spectrogram, spatial positions stand for the frequency bins of the sine and cosine functions, same patterns occurring in different positions may refer to completely different information; Second, CNNs exploit spatial locality by limiting the connectivity between neurons to adjacent regions, ending up small receptive field. Although this problem can be alleviated by pooling or cascading more layers, additional computations are introduced. It thus remains unclear what neural network architecture can capture the suitable inductive bias for signals in k-space.

In this paper, we propose a novel Transformer-based architecture for MRI reconstruction, termed as K-space Transformer, that enables to learn inductive bias in k-space. Inspired by the implicit representation [1,9], we treat the coordinates of unsampled k-space signal as inputs to the Transformer Decoder, iteratively query information from those sampled signals, enabling to model relationships between any frequency bins beyond the regular spatial grids. To summarise, our main contributions are three-fold: First, we introduce a novel hierarchical Transformer architecture that aims to learn suitable inductive bias for directly processing signals in k-space. Second, we integrate image domain refinement into the transformer module, and prove to be complementary to the pure k-space reconstruction; Third, we validate the effectiveness of K-space Transformer on two public datasets, demonstrating comparable or superior performance than previous state-of-the-art approaches.

2 Method

2.1 Problem Formulation

In this paper, we consider the problem of fast MRI reconstruction, with the goal of learning a function that maps the under-sampled spectrogram to high-quality MR images:

\[ I = \Phi(x_s; \Theta) \]  

(1)

where \( I \), \( x_s \) refers to the output MR image, and the input under-sampled spectrogram respectively, \( \Theta \) denotes the set of learnable parameters.

Unlike existing work on MRI reconstruction that generally exploit ConvNets in image or k-space, we adopt the implicit representation, and propose a novel Transformer-based architecture for learning the inductive bias in k-space. In the following sections, we start by introducing the fundamental building blocks, namely, the Encoder module (\( \Phi_{ENC} \)) that learns a compact feature representation from sampled frequency bins; and Decoder module (\( \Phi_{DEC} \)) that reconstructs


MRI by alternating completion in k-space and refinement in image domain. After that, we describe in detail how these blocks are used for constructing an hierarchical model, that strives a balance on the performance and computational complexity trade-off.

2.2 Encoder (Φ_{ENC})

Here, the goal is to compute a compact feature representation for the sampled frequency bins in k-space. In detail, given a set of n sampled points on spectrogram, i.e. \( \{s_1, \ldots, s_n\} \), with \( s_i = (m_i, p_i) \) refers to a combination of measured value \( m_i \in \mathbb{R}^2 \) and 2D spatial positions \( p_i \in \mathbb{R}^2 \). Note that, \( n \) is usually a small subset of all points on a spectrogram \( (\mathbb{R}^{W \times H}) \). As a tokenisation procedure that converts the sampled points into a vector sequence, \( V = \{v_1, v_2, \ldots, v_n\} \in \mathbb{R}^{n \times d} \), with

\[
v_i = \Phi_{\text{Tokenize}}(s_i) = \text{MLPs}(m_i) + \text{PE}(p_i)
\]

where one Multilayer Perceptrons (MLPs) are applied to the measured value, and \( \text{PE} \) refers to positional encodings with sine and cosine functions.

As shown in Figure 1 (a), the Encoder includes multiple standard transformer encoder layers, consisting of a multi-head self-attention module (MHSA), a feed forward network (FFN) and residual connections. Through self-attention module, the global dependency between each sampled point are captured, and the FFN further enriches the feature representation.

\[
O = \Phi_{\text{ENC}}(V) \in \mathbb{R}^{n \times d}
\]

where \( O \) denotes the output from encoder, with same dimensions as input sequence vectors. For more details, we would refer the readers to the original Transformer paper [14].
### 2.3 Decoder (Φ_{DEC})

In this section, we describe the decoding procedure that maps the encoded representation to MR images. We adopt an hybrid architecture that alternates between k-space decoding, and refinement in image domain, shown in Figure 1 (b).

**Decoding in k-space.** We adopt standard Transformer decoder layers, where the Key and Value are generated by applying two different linear transformations on encoder’s outputs, and the normalised positional coordinates for desired output (\(\mathbb{R}^W \times H\)) are encoded as Query:

\[
q_i = W^q \cdot \text{PE}(p_i) \quad \forall p_i \in \mathbb{R}^{W \times H}
\]

\[
k_i, v_i = W^k \cdot O_i, W^v \cdot O_i \quad \forall i \in [1, n]
\]

\[
S = g(\psi_{k-dec}(Q, K, V)) \in \mathbb{R}^{W \times H \times 2}
\]

where \(\text{PE}\) refers to positional encodings with sine and cosine functions, \(q_i, k_i, v_i\) refer to \(i\)-th vector of \(Q, K, V\) respectively, and \(S\) denotes the predicted spectrogram by applying a prediction layer \(g(\cdot)\) on the output from transformer decoders \(\psi_{k-dec}(\cdot)\), consisting of multi-head self-attention (MHSA), multi-head cross-attention (MHCA), a feed-forward network (FFN), and residual connections. Note that, the resolution of the reconstructed spectrogram can be controlled by the number of query vectors, i.e. the number of positional coordinates.

Thanks to the flexibility of Transformer decoder layers, dependencies between all frequency bins on spectrogram can be effectively captured, thus enable to learn the inductive bias in k-space. However, one issue is, the point-wise update in k-space will globally affects all pixels in image domain, disregarding the spatial bias in MRI, that is, structures tend to be roughly aligned. To resolve such issue, we alternate the processing in k-space and image domain.

**Refinement in Image Domain.** Here, we convert the reconstructed spectrogram into image domain with differentiable inverse Fast Fourier Transformer (FFT), and process the images with standard convolution layers \(\psi_{i-refine}(\cdot)\):

\[
S = T(\psi_{i-refine}(T^{-1}(S)))
\]

where \(T\) and \(T^{-1}\) refer to FFT and inverse FFT respectively. To avoid notation abuse, we also denote the output as \(S \in \mathbb{R}^{W \times H \times 2}\).

**Discussion:** At an high level, instead of treating the spectrogram as a look-up table, i.e. values on regular grids, we adopt the implicit representation, learning a continuous function that maps the input positional coordinates to values, conditioned on the visible k-space samples. Such representation is highly flexible, as it not only enables to explicitly model the dependencies between different frequency bins, but also support to output spectrogram of different resolutions by simply varying the density of query coordinates.
2.4 Hierarchical Decoder

With the basic building blocks introduced, we construct an hierarchical decoder, to strive a balance between the computational cost and performance trade-off. In particular, we adopt a low-resolution (LR) decoder and a high-resolution (HR) decoder to guide the reconstruction with progressive resolutions, and tailor them for different purposes:

$$\Phi_{\text{DEC}}(\cdot) = \Phi_{\text{HRD}} \circ \Phi_{\text{LRD}}(\cdot)$$

$\Phi_{\text{LRD}}$ aims to reconstruct the spectrogram on a lower resolution, focusing on the overall anatomical structure. As shown in Figure 1 (c), this is simply implemented by adapting the normalised positional coordinate $p_i$ in Equation 4 to a down-sampled spectrogram. Here, we do not use image domain refinement, and only implement the LR decoder with standard Transformer decoder layers.

The HR decoder aims to fill the texture details based on the LR output. We accordingly map the tokenised HR coordinates (Equation 4) to queries, and treat the two projections of LR decoder output as key and value. To alleviate the computational cost brought by self-attention module, we simplify each decoder layer in HR decoder with only cross-attention module left.

Discussion: As for computational complexity, when decoding the $m$-length sequence from the $n$-length sampled points with a standard Transformer decoder, the complexity would be $O(mn)$ for MHCA and $O(m^2)$ for MHSA, which is prohibitively expensive when $m$ is very large. However, in our case, by constructing the hierarchical model, the computation complexity is reduced to $O(ml + n + l^2)$, as the MIHA can simply be done in LR decoder, drastically cutting the memory consumption, when $l$ is set to be much smaller than $m$.

3 Experiment

3.1 Dataset & Metrics

We employ two datasets for evaluation, namely, OASIS and fastMRI. OASIS [8] contains 3,398 single-coil T1-weighted brain MRI volumes, each has 175 slices. We split them into train and test sets with ratio 3:1 and uniformly draw 10% from them. Additionally, fastMRI has 1,172 single-coil PD-weighted knee MRI volumes, each includes 35 slices. We use 973 for training, 199 for testing and uniformly draw 25% slices from both.

Following conventional evaluation, we experiment on five undersampling settings: Gaussian 2D sampling with 10×, 5× and 2.5× accelerations; Uniformly 1D sampling with 5× and 2.5× accelerations. Peak signal to noise ratio (PSNR) and structural index similarity (SSIM) are used as metrics.

3.2 Implementation Details

Our K-Space Transformer is implemented with 4 Encoder layers, 4 Low-Resolution Decoder layers and 6 High-Resolution Decoder layers. Following previous works [2,4,5,12,13],
we divide the complex MR signals into real and imaginary channels and enforce $\ell_2$ loss in image domain. Besides, to accelerate training, we also apply deep supervision on each layer of decoders. We adopt AdamW optimizer and cosine annealing schedule with initial learning rate of $5 \times 10^{-4}$. Our implement code will be available in https://github.com/zhaoziheng/K-Space-Transformer.

### 3.3 Comparison to State-of-the-art

We compare to 6 representative methods, including UNet [11] for image domain reconstruction and k-space reconstruction, Deep ADMM Net [16], D5C5 [12], and previous state-of-the-art OUCR [4].

As shown in Table 1, our method achieves the best results under all 1D settings. On OASIS, we improved the PSNR from 29.52 dB to 31.50 dB with 5× acceleration, and for 2.5× sampling, our improvement is even enlarged to 2.57 dB. On fastMRI, we outperform the best baseline slightly, however we also notice that the top approaches perform quite closely, we conjecture this is because the dataset is quite noisy, which may lead to an unprecise supervision for training and evaluation. In addition, it’s noticeable that K-UNet leads to poorer performance compared with UNet, showing that naïve CNN-based methods do not give suitable inductive bias for k-space reconstruction. On 2D settings (Table 2), our method either outperforms or achieve competitive results to the previous state-of-the-part approaches.

#### Table 1: Quantitative comparison on uniform 1D sampling pattern

| Method | OASIS | fastMRI |
|--------|-------|---------|
|       | $5 \times$ | $2.5 \times$ | $5 \times$ | $2.5 \times$ |
|       | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| K-UNet | 24.53 | 0.7808 | 26.50 | 0.8483 | 24.68 | 0.6058 | 26.99 | 0.7223 |
| UNet   | 27.53 | 0.8895 | 29.05 | 0.8988 | 26.72 | 0.6835 | 29.81 | 0.7983 |
| Deep ADMM | 26.03 | 0.7202 | 29.95 | 0.8166 | 24.43 | 0.6172 | 29.84 | 0.7838 |
| D5C5   | 27.66 | 0.8895 | 32.48 | 0.9615 | 26.87 | 0.6793 | 30.64 | 0.8126 |
| OUCR   | 29.52 | 0.9310 | 33.75 | 0.9709 | 27.71 | 0.7009 | 30.95 | 0.8168 |
| Ours   | 31.50 | 0.9528 | 36.32 | 0.9848 | 27.80 | 0.7078 | 31.04 | 0.8189 |

#### Table 2: Quantitative comparison on Gaussian 2D sampling pattern

| Method | OASIS | fastMRI |
|--------|-------|---------|
|       | $4 \times$ | $5 \times$ | $2.5 \times$ | $5 \times$ | $2.5 \times$ |
|       | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| K-UNet | 25.34 | 0.7666 | 28.76 | 0.8427 | 33.96 | 0.9288 | 27.15 | 0.6696 | 28.90 | 0.7710 |
| UNet   | 27.63 | 0.8437 | 31.27 | 0.9024 | 35.02 | 0.9388 | 28.33 | 0.7120 | 31.35 | 0.8283 |
| Deep ADMM | 26.17 | 0.7812 | 30.41 | 0.8122 | 33.94 | 0.8813 | 29.31 | 0.7515 | 30.97 | 0.8162 |
| D5C5   | 30.51 | 0.9241 | 35.91 | 0.9743 | 42.72 | 0.9927 | 29.78 | 0.7507 | 32.00 | 0.8399 |
| OUCR   | 31.61 | 0.9342 | 36.78 | 0.9797 | 42.80 | 0.9944 | 29.95 | 0.7522 | 32.04 | 0.8392 |
| Ours   | 32.58 | 0.9520 | 37.47 | 0.9823 | 43.68 | 0.9946 | 29.83 | 0.7534 | 31.97 | 0.8389 |

### Qualitative Results:

In Figure 2, we show some predictions from our proposed K-space Transformer. On OASIS, especially under the uniform sampling pattern, our method clearly retain more texture information and less structural loss compared with others, suggesting that our method is more robust for such
severe distortion in image domain. While on fastMRI, the groundtruth images are often noisy, which might explain why the top approaches perform similarly.

![Figure 2: Qualitative comparison of 5× acceleration on different sampling patterns. Brighter means higher error.](image)

### 3.4 Ablation Study

To investigate the effectiveness of hybrid learning and the hierarchical structure, we remove the image domain refinement module (RM), and the LR decoder (LRD) sequentially on OASIS dataset. As shown in Table 3, without the
refinement module, the reconstruction quality drops significantly, showing the image domain refinement can indeed provide supplementary information for k-space reconstruction. Nevertheless, our method still achieves competitive results and outperforms K-UNet by a large margin (refer to Table 1 and 2), which justifies Transformer’s superiority for k-space reconstruction. In addition, removing the LR decoder further leads to an obvious performance degradation, implying that our hierarchical design also plays an essential role in guaranteeing an excellent reconstruction quality with a moderate computational cost.

| Modules | Gaussian 2D 5× | Uniform 1D 5× |
|---------|----------------|----------------|
|         | PSNR | SSIM   | PSNR | SSIM   |
| ✓       | 37.47 | 0.9823 | 31.50 | 0.9528 |
| ✓       | 32.97 | 0.9337 | 27.52 | 0.8876 |
| ×       | 28.53 | 0.8416 | 21.31 | 0.6114 |

Table 3: Quantitive ablation study on critical designs.

We visualize some intermediate results of K-space Transformer in Figure 3. We observe the LR decoder capable to reconstruct a basic outline of the structure, equally the low-mid frequency bands of the spectrogram. Based on the upsampled rough result, the HR decoder fills in the texture details, i.e. completing higher frequency bands. The coarse-to-fine progress is consistent with our assumption in Section 2.4. Compared to standard HR decoder, those without refinement module contain more noise over the structures and background, which is easier to remove via spatial convolution. It indicates the inductive bias of image domain are complementary to k-space.

**Fig. 3:** Intermediate results of K-space Transformer: LR denotes the reconstruction output of LR decoder; Upsampled denotes the up-sampled result; HR and HR(wo RM) refer to the output of HR decoder and that without refinement module.
3.5 Visualisation

We visualize the attention maps of our encoder and decoder according to:

$$\text{attn} = \text{softmax}(QK^T/\sqrt{d})$$

where $Q$ and $K$ refer to the Query and Key indicated in section 2.2 and 2.3. As shown in Figure 4, we observe that both the global and local interactions between frequency bins have been well exploited through the attention mechanism, as there are some heads clearly scanning the spectrogram globally, and others focusing on specific regions of interests, implying that the model can indeed capture inductive bias in k-space.

Fig. 4: Attention maps from the encoder and decoder under different masks. The upper rows come from the self-attention modules in encoder, while the lower rows are from the cross attention modules in high-resolution decoder. We point out coordinates of the selected sampled frequencies with red arrow, and coordinates of unsampled frequencies with blue arrow.

4 Conclusion

We propose K-space Transformer, a novel Transformer-based architecture for fast MRI reconstruction. It exploits implicit representation, and learns the global dependencies between frequencies in k-space. Specifically, we design a flexible hierarchical decoder that strives to produce high quality reconstruction with
moderate computational cost. Extensive experiments show that the proposed K-space Transformer outperforms existing state-of-the-art methods or achieves comparable performance on two public datasets, demonstrating promising results of Transformer in MRI reconstruction and k-space learning. Future studies can be conducted to adapt it to multi-coil imaging, k-space super-resolution and other potentially relevant directions.

References

1. Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8628–8638, 2021.
2. Tianming Du, Honggang Zhang, Yuemeng Li, Stephen Pickup, Mark Rosen, Rong Zhou, Hee Kwon Song, and Yong Fun. Adaptive convolutional neural networks for accelerating magnetic resonance imaging via k-space data interpolation. Medical Image Analysis, 72:102098, 2021.
3. Taejoon Eo, Yohan Jun, Taeseong Kim, Jinseong Jang, Ho-Joon Lee, and Dosik Hwang. Kiki-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. Magnetic resonance in medicine, 80(5):2188–2201, 2018.
4. Pengfei Guo, Jeya Maria Jose Valanarasu, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M Patel. Over-and-under complete convolutional rnn for mri reconstruction. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 13–23. Springer, 2021.
5. Yoseo Han, Leonard Sunwoo, and Jong Chul Ye. K-space deep learning for accelerated mri. IEEE transactions on medical imaging, 39(2):377–386, 2019.
6. Kieren Grant Hollingsworth. Reducing acquisition time in clinical mri by data undersampling and compressed sensing reconstruction. Physics in Medicine & Biology, 60(21):R297, 2015.
7. Dongwook Lee, Jaegun Yoo, and Jong Chul Ye. Deep residual learning for compressed sensing mri. In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pages 15–18, 2017.
8. Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience, 19(9):1498–1507, 2007.
9. Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In European conference on computer vision, pages 405–421. Springer, 2020.
10. Osvald Nitski, Sayan Nag, Chris McIntosh, and Bo Wang. Cdf-net: Cross-domain fusion network for accelerated mri reconstruction. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 421–430. Springer, 2020.
11. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.
12. Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2017.

13. Roberto Souza, R Marc Lebel, and Richard Frayne. A hybrid, dual domain, cascade of convolutional neural networks for magnetic resonance image reconstruction. In *International Conference on Medical Imaging with Deep Learning*, pages 437–446. PMLR, 2019.

14. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

15. Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al. Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2017.

16. Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Admm-csnet: A deep learning approach for image compressive sensing. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):521–538, 2018.

17. Simiao Yu, Hao Dong, Guang Yang, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, David Firmin, et al. Deep de-aliasing for fast compressive sensing mri. *arXiv preprint arXiv:1705.07137*, 2017.

18. Gushan Zeng, Yi Guo, Jiaying Zhan, Zi Wang, Zongying Lai, Xiaofeng Du, Xiaobo Qu, and Di Guo. A review on deep learning mri reconstruction without fully sampled k-space. *BMC Medical Imaging*, 21(1):1–11, 2021.

19. Bo Zhou and S Kevin Zhou. Dudornet: learning a dual-domain recurrent network for fast mri reconstruction with deep t1 prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4282, 2020.