Towards Architecture-Agnostic Un-trained Network Priors for Image Reconstruction with Frequency Regularization

Yilin Liu[1], Yunkui Pang[1], Jiang Li[1], Yong Chen[2], and Pew-Thian Yap[1]

[1] Computer Science, University of North Carolina at Chapel Hill
[2] Radiology, Case Western Reserve University

1 Natural image experiments

1.1 Inpainting

To the best of our knowledge, our method is the first attempt that addresses challenges related to architecture, overfitting and runtime simultaneously.

To compare with prior DIP methods, we 1) first employ the network configurations that work best in their respective settings [3, 13], i.e., the architectures used are different in different methods, as shown in Fig. 1 and Fig. 2, we then 2) use an underperforming architecture for all competing methods, as shown in Fig. 3.

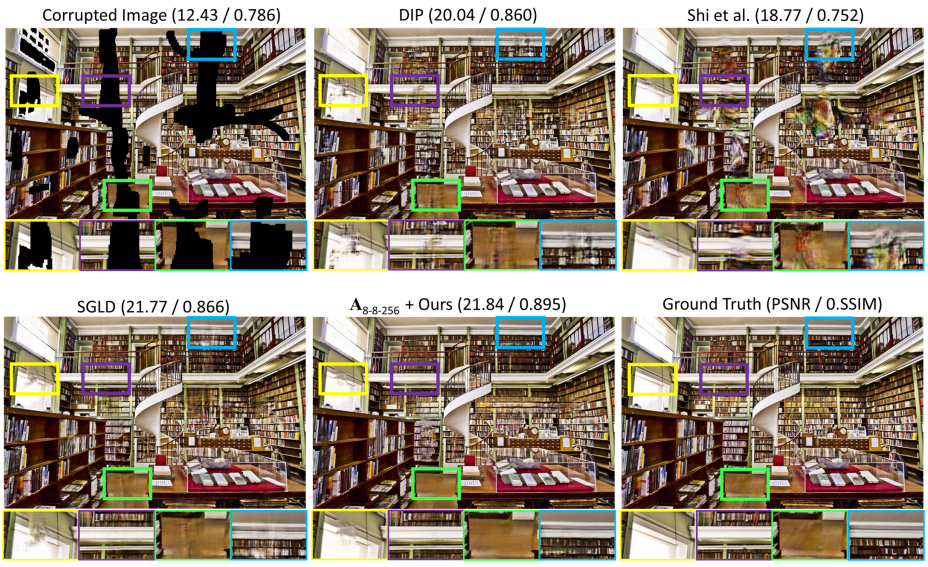

Fig. 1: Qualitative comparisons with previous DIP methods on inpainting.

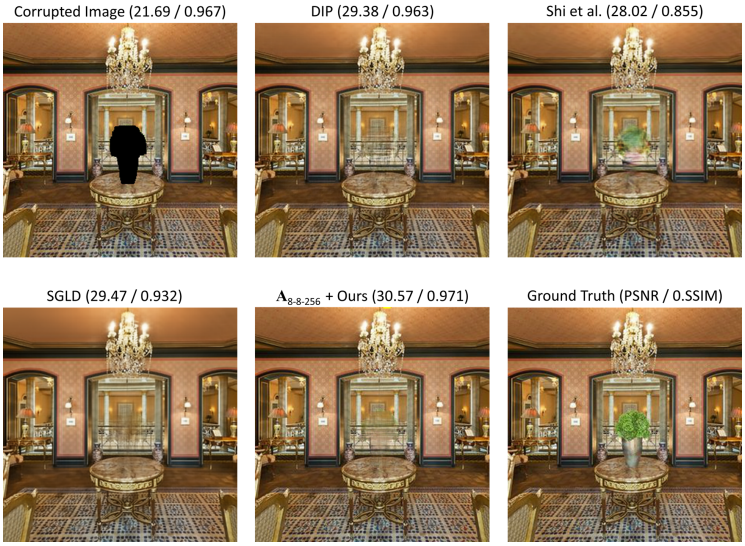

Fig. 2: Qualitative comparisons with previous DIP methods.

Fig. 3: When unsuitable architectures are used, e.g., UNet w. deconvolutions, ResNet w/o. upsampling, SGLD [3] fails to perform restoration. This confirms the importance of architectural decisions in DIP, and that previous methods do not address the architecture-related challenges. Note that deconvolutions have been reported to be not suitable for DIPs [7, 10]. Similarly for ResNet, which does not have any upsampling layers [10].

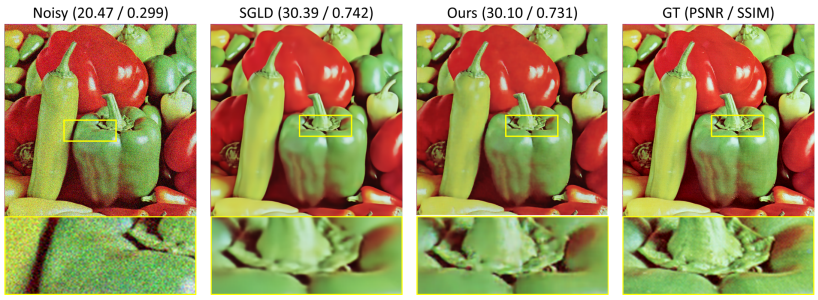

Fig. 4: Qualitative comparisons with previous DIP methods on denoising. Ours trades off the metrics for sharpness.

Fig. 5: Denoising results. The base network used in each competing method is **replaced by an underperforming architecture, i.e., $\mathbf{A_{2-256}}$.**

1.2 Denoising

We first show the results when each method operates in its original setting (Fig. 4), and then evaluate them when an unsuitable architecture is used (Fig. 5).

Transformer. Besides CNN, we show here the result on Swin U-Net [1], which consists of only Swin Transformer blocks and skip connections, i.e., no upsampling is involved. As noted in a recent study [10], the unlearnt upsampling is the driving

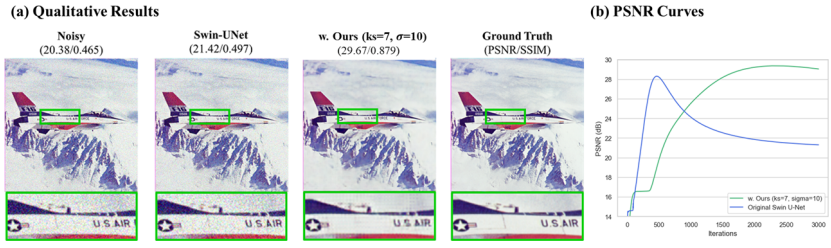

Fig. 6: Qualitative denoising results of a **transformer** [1]. Our method substantially alleviates the overfitting and enhances the peak PSNR.

force behind the spectral bias of DIP, and such transformers are more difficult to perform denoising. This finding assumes the white noise as the input. Here, we show that constraining the bandwidth of the noise input enables long-lasting denoising even in such models.

2 Influences of hyperparameters

As our methods contain several hyperparameters, we visualize their influences on the frequency control and hence the regularization effects over the output image in Fig. 7.

s and σ are associated with the Gaussian blur kernel applied on the noise input. The larger the s and/or σ is, the more high frequencies are removed from the noise input (i.e., the smoother the output is). M and β are for adjusting the attenuation extents of the Kaiser-based upsamplers. The higher the M and/or β is, the larger the attenuation of the high-frequency replica (i.e., the smoother the output is).

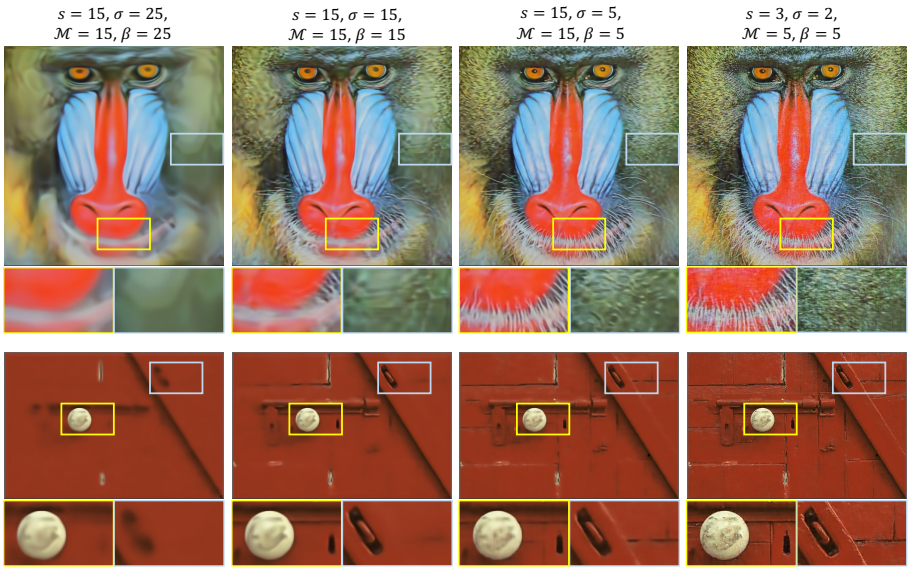

Fig. 7: Visualizations of the frequency control with our methods in denoising experiments.

3 Comparisons with early-stopping

To complement Table 5 in the main text, here we visualize in Fig. 8 that early stopping, even though prevents further performance decay, cannot fundamentally

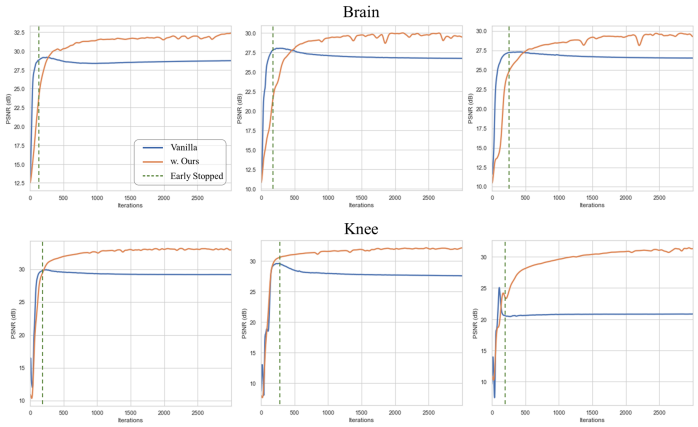

Fig. 8: Comparisons with self-validation-based early stopping. Although early stopping alleviates overfitting, it cannot fundamentally improve the underperforming architectures. Our methods can improve their peak performance while also mitigating overfitting.

improve the underperforming architectures. In other words, **early stopping could not cope with ill-designed architectures**.

4 Architectural influences

To better inform our method design, we 1) investigate the architectural influences in the context of MRI reconstruction, and also 2) validate our findings on image inpainting and denoising. Our results confirm that the reconstruction outcome is sensitive to basic architectural properties.

4.1 Crucial Architectural Elements

We first pinpointed the *core* architecture elements that have a critical impact on the performance.

Experimental setup i. Since a decoder is the minimum requirement for reconstruction, we experimented with two types of 7-layered decoder-only architectures, i.e., ConvDecoder [5] and Deep Decoder [6]. Experiments were performed on the 4× under-sampled multi-coil knee MRI from fastMRI database [9].

Upsampling (interpolation filter). Fig. 9(a) suggests an interesting result: removing the *unlearnt* upsampling, e.g., bilinear, leads to either failure or unstable results (see gray curves). Unlike transposed convolution, the unlearnt upsampler is essentially a *fixed* low-pass interpolation filter that attenuates the introduced high-frequency replica and also the signal. Frequency response of bilinear interpolation filter decays more rapidly than that of nearest neighbor as the frequency increases (Fig. 9 (b)), suggesting stronger attenuation and smoothing effects. Hence, bilinear

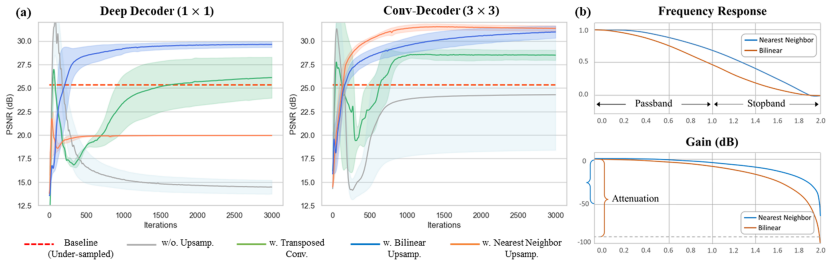

Fig. 9: Influences of architectural elements. Results averaged across three different widths.

upsampling typically biases the network towards generating smoother outputs, as prevalent in generative models [12]. Transposed convolutions, however, are not guaranteed to be low-passed as they are *learnable*. Due to the spectral bias of network layers, they may be low-passed during early training to still enable reconstruction, but the results could be unstable (green curves).

Convolutional layers. When the unlearnt upsampling operations are absent, ConvDecoder (3×3) still enables reconstruction while Deep Decoder (1×1) fails completely (Fig. 9(a)). A similar phenomenon is also reported in image denoising [2, 10]. This again be attributed to CNN's inherent spectral bias and suggests that the size of the kernel also matters, further corroborated in Tab. 2.

Discussion. Results of this pilot experiment suggest that the spatial kernels with low-pass characteristics, either learnable or unlearnt, are *crucial* to the success of untrained network priors. In particular, bilinear upsampling with a fixed low-pass filter produces more stable and better results (blue curves).

4.2 Depth, Width and Skip Connections

Here, we demonstrate that insights gained about the unlearnt upsampling can aid in understanding the connection between architectural characteristics and the reconstruction task.

Experimental setup ii. For this large-scale validation, we experimented with an isotropic encoder-decoder architecture used in the original DIP, i.e., equal width and kernel size for all layers throughout the network. Design choices are detailed in Tab. 1. Experiments were performed on the publicly available 4× under-sampled multi-coil knee MRI from fastMRI database [8].

Table 1: Test bed for studying the architectural influences of an encoder-decoder untrained networks.

Archi. Type	Depth (d)	# of Skips (s)	Width (w)	Kernel Size (k)
$\mathbf{A_{d-s-w-k}}$	{2-L, 3-L, 4-L, 5-L, 8-L}	{zero, half, full}	{32, 64, 128, 256}	$\{3 \times 3, 5 \times 5\}$

Table 2: Influences of typical architectural design choices in knee reconstruction. *Deeper* and/or *Narrower* architectures tend to perform better; skip connections influence the deep architectures more; larger kernels perform slightly better. $\mathbf{A_{8\text{-}full\text{-}32\text{-}3}}$ performs the best (in lime); $\mathbf{A_{2\text{-}full\text{-}256\text{-}3}}$ performs the worst (in red).

	Width (↓)											
	Archi.	PSNR	SSIM	Archi.	PSNR	SSIM	Archi.	PSNR	SSIM	Archi.	PSNR	SSIM
Depth (↑)	$\mathbf{A_{2\text{-}full\text{-}256\text{-}3}}$	26.67	0.530	$\mathbf{A_{2\text{-}full\text{-}128\text{-}3}}$	27.12	0.543	$\mathbf{A_{2\text{-}full\text{-}64\text{-}3}}$	27.70	0.583	$\mathbf{A_{2\text{-}full\text{-}32\text{-}3}}$	28.47	0.641
	$\mathbf{A_{3\text{-}full\text{-}256\text{-}3}}$	28.22	0.590	$\mathbf{A_{3\text{-}full\text{-}128\text{-}3}}$	28.59	0.605	$\mathbf{A_{3\text{-}full\text{-}64\text{-}3}}$	28.55	0.616	$\mathbf{A_{3\text{-}full\text{-}32\text{-}3}}$	29.25	0.660
	$\mathbf{A_{4\text{-}full\text{-}256\text{-}3}}$	28.68	0.617	$\mathbf{A_{4\text{-}full\text{-}128\text{-}3}}$	28.95	0.622	$\mathbf{A_{4\text{-}full\text{-}64\text{-}3}}$	28.87	0.624	$\mathbf{A_{4\text{-}full\text{-}32\text{-}3}}$	29.70	0.671
	$\mathbf{A_{5\text{-}full\text{-}256\text{-}3}}$	28.61	0.613	$\mathbf{A_{5\text{-}full\text{-}128\text{-}3}}$	28.87	0.615	$\mathbf{A_{5\text{-}full\text{-}64\text{-}3}}$	29.33	0.648	$\mathbf{A_{5\text{-}full\text{-}32\text{-}3}}$	29.81	0.680
	$\mathbf{A_{8\text{-}full\text{-}256\text{-}3}}$	28.98	0.625	$\mathbf{A_{8\text{-}full\text{-}128\text{-}3}}$	29.33	0.637	$\mathbf{A_{8\text{-}full\text{-}64\text{-}3}}$	29.45	0.651	$\mathbf{A_{8\text{-}full\text{-}32\text{-}3}}$	30.04	0.695
	Skip Connections (−)						Kernel Size (↑)					
	$\mathbf{A_{2\text{-}half\text{-}256\text{-}3}}$	26.91	0.535	$\mathbf{A_{2\text{-}zero\text{-}256\text{-}3}}$	26.83	0.535	$\mathbf{A_{2\text{-}full\text{-}256\text{-}3}}$	26.67	0.530	$\mathbf{A_{2\text{-}full\text{-}256\text{-}5}}$	26.98	0.550
	$\mathbf{A_{4\text{-}half\text{-}256\text{-}3}}$	28.55	0.621	$\mathbf{A_{4\text{-}zero\text{-}256\text{-}3}}$	27.54	0.697	$\mathbf{A_{5\text{-}full\text{-}256\text{-}3}}$	28.61	0.613	$\mathbf{A_{5\text{-}full\text{-}256\text{-}5}}$	28.82	0.624
	$\mathbf{A_{8\text{-}half\text{-}256\text{-}3}}$	29.12	0.669	$\mathbf{A_{8\text{-}zero\text{-}256\text{-}3}}$	28.51	0.609	$\mathbf{A_{8\text{-}full\text{-}256\text{-}3}}$	28.98	0.625	$\mathbf{A_{8\text{-}full\text{-}256\text{-}5}}$	29.12	0.634

What do deeper and narrower architectures produce? (Tab. 2). Theoretically, as the number of layers (depth) or channels (width) increases, the ability of the network to learn arbitrarily high frequencies (details, noise) is typically increased [11]. While this is true for width, we have found that the effect on depth turns out to be attenuated by unlearnt upsampling. As evidenced in Fig. 10, deeper architectures typically generate smoother images, exhibiting a stronger preference for low-frequency information, whereas shallower counterparts, *even though they have fewer parameters*, are more susceptible to noise and overfitting (red arrows). This is more evident when comparing the *same* architectures with just different upsamplers, where the architectures with bilinear upsampling (stronger attenuation) are less prone to overfitting than the ones using nearest neighbor (NN) upsampling (cyan vs. blue). Hence, it is not merely the number of parameters but the architectural characteristics promoting low frequencies that seem to be the primary reason for the high performance. Note that all these results are only achievable when *unlearnt* upsampling is involved (gray dashed curves).

Skip connections. Deep architectures with zero skip connection converge more slowly and may lead to over-smoothing as shown in Fig. 10 (red curves). Skip connections greatly alleviate this issue and introduce more details (cyan curves), which we speculate could be due to the "reduced effective up-sampling rate". Yet, excessive skip connections make a deep architecture behave similarly as a

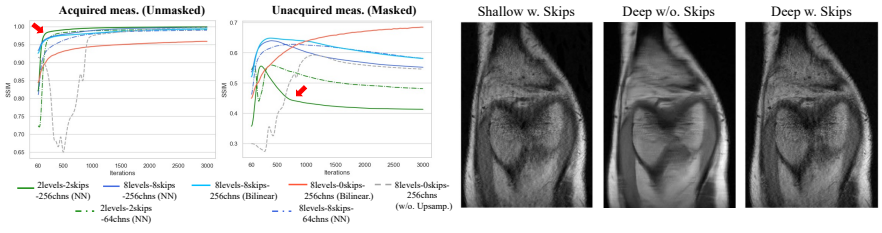

Fig. 10: Generalizability of different architectures on the masked regions.

shallower one, generating more noise (Fig. 10 right). Overall, they exert a greater influence on deeper architectures ($\mathbf{A_{8\text{-}zero}} < \mathbf{A_{8\text{-}full}} < \mathbf{A_{8\text{-}half}}$) compared to shallower ones ($\mathbf{A_{2\text{-}full}} \approx \mathbf{A_{2\text{-}zero}}$).

4.3 Validations on image denoising and inpainting.

We reaffirmed our observations above on image denoising and inpainting, as shown in Fig. 11, Fig. 13, Fig. 14 and Fig. 12.

We argue that the understanding of the upsampling and its interactions with other architectural elements can help explain why deeper networks with fewer skip connections converge more slowly, generate smoother outputs and are less prone to overfitting. Concretely, the upsampling operation inserted in-between the decoder layer slows down the generation of high frequencies required for transforming the lower-resolution feature maps into the higher-resolution target image, primarily due to its role as a fixed low-pass filter. As the network depth increases, the degree of smoothness increases (Fig. 13). Skip connections notably accelerate the convergence (Fig. 12) and ameliorate the over-smoothing issue, likely due to the reduced "effective" upsampling rate. All these observations are consistent with our MRI experiments in Sec. 2.

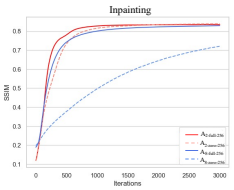

Fig. 12: Deep architectures with zero skip connection converge more slowly, i.e., $\mathbf{A_{8\text{-}zero\text{-}256}}$

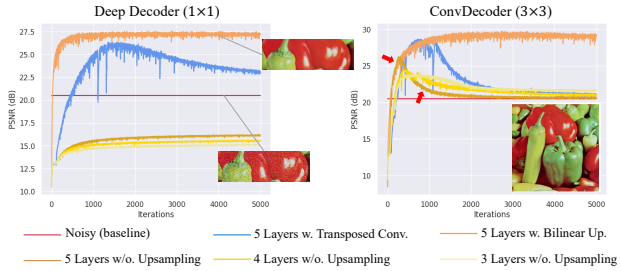

Fig. 11: Denoising experiments. (**Left**) In non-convolutional networks, removing the upsampling hampers the denoising capability, which cannot be compensated by merely adjusting the network to be more under-parameterized. Transposed convolutions result in a more rapid decline in performance than bilinear upsampling. (**Right**) Convolutional layers *alone* exhibit certain denoising effects but necessitate early stopping. The showcased image is from the classic dataset Set9 [4].

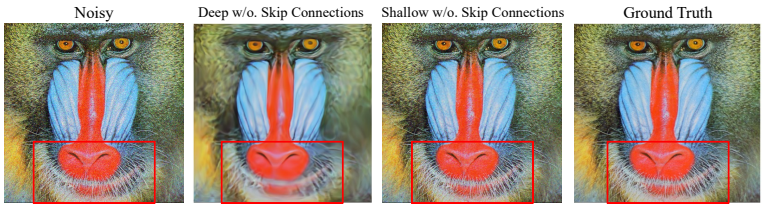

Fig. 13: Denoising experiments. Deeper architectures with few or no skip connections tend to generate smoother outputs compared to the shallower ones.

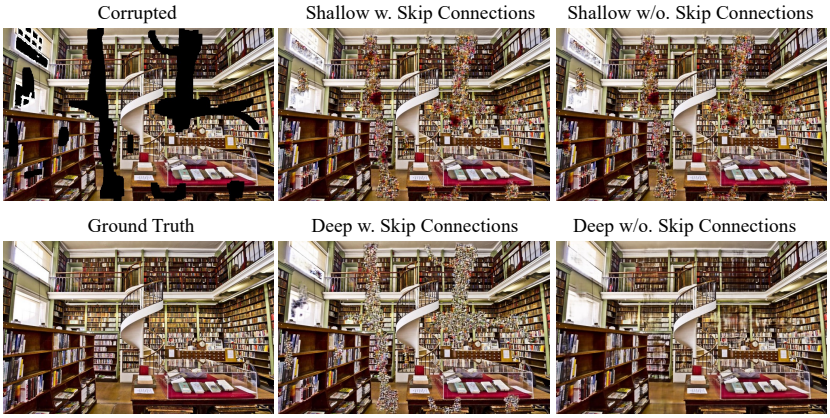

Fig. 14: Inpainting experiments. Deeper architectures with few or no skip connections tend to generate smoother predictions for the masked regions than the shallower architectures. Skip connections make deep architectures perform similarly as the shallower ones.

5 Comparisons with `ZS-SSL-UNet`

In the main paper, we have included the results of the ResNet version of `ZS-SSL` [14]. Here, we constructed a UNet variant of it, dubbed `ZS-SSL-UNet`. As shown in Tab. 3, the architecture type impacts not only DIP but also deep unrolling networks, and potentially a broader area, which worth future investigations.

Table 3: Quantitative evaluations. Runtime: mean ± std mins per slice.

	fastMRI Brain		fastMRI Knee		Stanford FSE		Runtime
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	Avg
`ZS-SSL-ResNet` [14]	34.39	0.878	32.00	0.773	31.74	0.805	45.5 ± 11.8 mins
`ZS-SSL-UNet`*	25.70	0.670	27.79	0.703	26.73	0.674	97.5 ± 41.2 mins
Ours	33.10	0.874	32.07	0.781	31.30	0.800	4.4 ± 0.4 mins

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
2. Chakrabarty, P., Maji, S.: The spectral bias of the deep image prior. arXiv preprint arXiv:1912.08905 (2019)
3. Cheng, Z., Gadelha, M., Maji, S., Sheldon, D.: A bayesian perspective on the deep image prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5443–5451 (2019)
4. Dabov, K., Foi, A., Egiazarian, K.: Video denoising by sparse 3d transform-domain collaborative filtering [c]. In: Proc. 15th European Signal Processing Conference. vol. 1, p. 7 (2007)
5. Darestani, M.Z., Heckel, R.: Accelerated mri with un-trained neural networks. IEEE Transactions on Computational Imaging **7**, 724–733 (2021)
6. Heckel, R., Hand, P.: Deep decoder: Concise image representations from untrained non-convolutional networks. arXiv preprint arXiv:1810.03982 (2018)
7. Heckel, R., Soltanolkotabi, M.: Denoising and regularization via exploiting the structural bias of convolutional generators. International Conference on Representation Learning (2020)
8. Knoll, F., Murrell, T., Sriram, A., Yakubova, N., Zbontar, J., Rabbat, M., Defazio, A., Muckley, M.J., Sodickson, D.K., Zitnick, C.L., et al.: Advancing machine learning for mr image reconstruction with an open competition: Overview of the 2019 fastmri challenge. Magnetic resonance in medicine **84**(6), 3054–3070 (2020)
9. Knoll, F., Zbontar, J., Sriram, A., Muckley, M.J., Bruno, M., Defazio, A., Parente, M., Geras, K.J., Katsnelson, J., Chandarana, H., et al.: fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. Radiology: Artificial Intelligence **2**(1), e190007 (2020)
10. Liu, Y., Li, J., Pang, Y., Nie, D., Yap, P.T.: The devil is in the upsampling: Architectural decisions made simpler for denoising with deep image prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12408–12417 (2023)
11. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: International Conference on Machine Learning. pp. 5301–5310. PMLR (2019)
12. Schwarz, K., Liao, Y., Geiger, A.: On the frequency bias of generative models. Advances in Neural Information Processing Systems **34**, 18126–18136 (2021)
13. Shi, Z., Mettes, P., Maji, S., Snoek, C.G.: On measuring and controlling the spectral bias of the deep image prior. International Journal of Computer Vision **130**(4), 885–908 (2022)
14. Yaman, B., Hosseini, S.A.H., Akçakaya, M.: Zero-shot self-supervised learning for mri reconstruction. International Conference on Learning Representations (2022)