Spectral Saturation in Deep Equilibrium Models: A Comparative Analysis of Conservative Residual Refinement on Texture versus Structure

Daniel Woodford
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA
dfw5416@psu.edu

Joshua Wufsus
College of Engineering
The Pennsylvania State University
University Park, PA
jpw6234@psu.edu

***Abstract*—Deep Equilibrium Models (DEQs) are a relatively new (2019) set of models that have proven to be useful for Implicit Neural Representations (INRs). While investigating the TorchDEQ [1], of which we base our approach upon, there are many examples that show a convergence floor that does not fully capture some frequency details of the original image. To mitigate this, we are proposing and evaluating a Conservative Residual Refinement (CRR), which pairs a DEQ with a shallow refinement network, aimed at increasing PSNR (peak signal-to-noise ratio) beyond the convergence point of the baseline DEQ. However, ensuring stability within the model remains at the forefront of importance, hence the conservative nature of the refiner. While stable, we found a dichotemy in performance. On texture-rich, continuously-signaled images, this method provides significant PSNR gains (+0.18 dB), but the performance degrades on structure-dominant images. This paper investigates why that may be the case, and suggests that CRR highlights a fundamental trade-off when attempting additive residual learning for INRs.**

I. Introduction

Implicit Neural Representations (INRs) have been a recent revolution in parameterizing continuous signals, yet they remain difficult to optimize. Deep Equilibrium models (DEQs) seem to be a promising solution, by replacing explicit layers with a fixed-point iteration $z^* = f_\theta(z^*, x)$. This allows for an "infinite depth" without the infinite memory consumption that a layered model would have with this depth. As noted by Gilton et al., this is what makes DEQs consistently be able to make improvements in reconstruction accuracy for inverse imaging problems compared to fixed-depth models. [1]

However, the stability constraints currently required to consistently solve for these fixed points (e.g., Lipschitz, Jacobian regularization) often add a "spectral bias", which essentially low-pass filters the output. We introduce **Conservative Residual Refinement** as a potential fix. Unlike joint training methods which may lead to destabilized convergence, and thus a failed equilibrium, CRR treats the converged DEQ as a sort of "base learner", and then trains a shallow network to predict the residual error $r = y_{gt} - y_{deq}$.

Our contributions are as follows:

1) **Method:** We use a two-stage training methodology using "Conservative Initialization" ($\sigma \approx 10^{-5}$) to prevent downgraded inference with the already-trained fixed point.
2) **Frequency Analysis:** We demonstrate that our CRR is not always beneficial. It is able to synthesize textures from high-frequency signals, but fails to generate the proper amount of noise without these signals.
3) **Synthesis:** We integrate the findings from Deep Equilibrium Object Detection (DEOD) [2] and Deep Equilibrium Models for Snapshot Compressive Imaging [3] to explain why refining utilizing specific pixel values cannot correct structural errors within the reconstruction, which pushes for a move toward semantic or structural query refinement.

II. Related Work

1) **Deep Equilibrium Models in Imaging** Utilizing DEQs for imaging tasks is well-documented at this point. Gilton et al. demonstrated that DEQs can be very useful for inverse imaging problems such as MRI reconstruction and deblurring. They argue that the traditional method, unrolled networks, are limited by the fixed and small number of iterations. DEQs on the other hand allow for an infinite number of iterations, leading to consistent improvements in reconstruction accruacy. Our work builds on this by enhancing this "infinte depth" result with a finite post-processor. [1]
2) **Stability and Regularization** Zhao et al. explores the advantages of DEQs for Snapshot Compressive Imaging (SCI) [3]. They argue that a primary advantage of DEQs is the combination of regularization and stable convergence. The implicit layer is able to naturally filter out inconsistencies. Our approach risks missing out on this implicit regularization, which goes against this purported core strength of DEQs.
3) **Structural Refinement vs. Pixel Refinement** Wang et al. (DEOD) [2] focuses on refinement for object detection. They suggest that explicit queries with the ability to encode information such as locations and categories is important for a powerful refinement. The

alternative is using just pixel intensities, which cannot correct structural errors in the reconstructed image. In DEOD, the equilibrium solver works by refining these structural queries directly. Our proposed shallow refiner is in direct contrast, since it lacks semantic awareness and operates only on the residual errors of pixel intensities. We believe encoding object-level information for general image reconstructing defeats the purpose of a generalizable approach. Designating locations and categories of objects seems to make much more sense in the context of object detection that image reconstruction.

III. Methodology

Mathematically, a layer function is represented by

$$f(X) = \lambda(\alpha X + \beta) \tag{1}$$

where α represents the vector of neuron weights assigned in the training process, β represents a scalar bias term, and X represents the input vector of neuron values. λ represents the activation function, which is a function applied similarly to each neuron within the layer. When multiple layer functions are applied to an input, X, consecutively, an artificial neural network is created. In practice, the output $y = f(x)$ is passed as input to the next neural function (i.e., $f_2(y)$).

Typically, an increase in layers is associated with better accuracy, at the tradeoff of increased computational cost and increased memory usage. This is because the larger a model is, the greater the capacity the model has to learn complex patterns through a deeper hierarchical model state.

One way to address this is by building a neural network which uses a single hidden layer in practice, but simulates the presence of infinite layers. This subtype of artifical neural networks are called Deep Equilibrium Networks.

In order to simulate infinite recursion, Deep Equilibrium Networks (DEQs) require a fixed point as the weight vector. A fixed point is an area in the domain of a layer function that returns the same input as the output at all times.

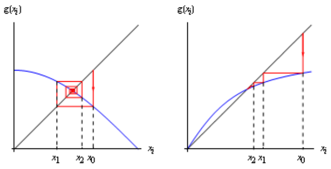

Fig. 1. Fixed Point Iterations

Specifically, the goal of fixed point solving in the training of a neural network is to find a point in the domain, z^*, such that the output is equal to the input (i.e., $z^* = f(z^*)$). Figure 1 visualizes the fixed point solving process. In this image, the red line can be interpreted as the step in the direction of the gradient. For example, if the gradient points directly to the left at any given moment, the line would move horizontally to the left for a given step size. This gradient descent process is then repeated until the system reaches a point where the gradient is 0 and the system has achieved convergence. If the gradient of a layer function, ∇f is 0, this means all partial derivatives are 0, and the output of the layer function will be equal to the input. This desirable point where the gradient reaches zero is the fixed point. A direct consequence of using a fixed point is that the output of one layer, and the output of 10 layers will both return the original input X. In this way, fixed point solving enables an implied infinite layer depth at constant runtime and memory usage.

Our unique methodology involves a multistep training process aimed at boosting the performance of a baseline DEQ. The first step is to train a DEQ until convergence is reached. That is, finding a point, z^*, where the system converges to return the same input. This baseline DEQ is expected to do well with low frequency data and locks in the global structure.

Once this network is trained, it is effectively “frozen” in place, no further training is done. The next step is adding a shallow refiner network, which uses both the input X and the residuals of the first DEQ as input. Explicitly, the formula is represented by

$$z^* = \sigma(Wz^* + u(x) + b)$$

We use different symbols to differentiate between traditional layer functions and a Deep Equilibrium layer function, W represents the Weight matrix now, and z^* represents the fixed input point. The important part of this piece is the initialization of the final layer of the refiner. By initializing the weights and bias to near 0 terms, we ensure the training process of the refiner starts off where the frozen DEQ left off. This allows the refiner to learn more complex imagery while also maintaining the ability to predict low-frequency data.

One great benefit of this approach is that the output is the best of both worlds. By starting the refiner training process where the baseline DEQ left off, the compounded output is guaranteed to be at least as good as the baseline DEQ. When high frequency, complex data is passed into our system, the shallow refiner offers an ability to accurately process this beyond the ability of the baseline DEQ. When low frequency data is passed to our system, output is still guaranteed to be at least as good as the baseline DEQ, which thrives on this type of input.

IV. Reconstruction Analysis

A. Setup

We utilized two benchmark images from the *skimage* package, preprocessed to 512x512 resolution. These two images were chosen as they were the only ones tested by the parent paper, and they alone had differing characteristics to test our hypotheses on. These models are also only able to be trained on one image (to replicate), so there is no sort of extra PSNR gained by training on more.

B. Maintaining the Integrity of the Specifications

Hyperparameters were held constant between trials. 500 epochs for the baseline and 300 epochs for the refiner.

Fig. 2. Astronaut

Fig. 3. Photographer

C. Quantitative Results & Analysis

The table below showcases the Peak Signal-to-Noise Ratio (PSNR) at the end of training:

TABLE I
QUANTITATIVE COMPARISON OF RECONSTRUCTION FIDELITY (PSNR)

Image Signal	**PSNR Performance Metrics**		
	Stage 1: DEQ (dB)	***Stage 2: Refined (dB)***	***Net Improvement***
Astronaut	35.06	35.22	**+0.18 dB**
Photographer	35.59	35.58	-0.01 dB

[a] Bold text indicates successful improvement.

Our CRR was able to push the reconstructed PSNR for the astronaut by .18 dB, indicating that our approach was successfully able to enhance the learning of the DEQ. The photographer image reconstruction lost .01 db, perhaps due to rounding errors or other small issues since theoretically, the CRR should only improve or maintain image reconstruction. Later, we theorize as to why discrepancies in results exist between the two images. Furthermore, Fig. 4-5 below display the PSNR (dB) and Loss Curves on the image reconstruction of the conjoined DEQ+CRR on the astronaut image:

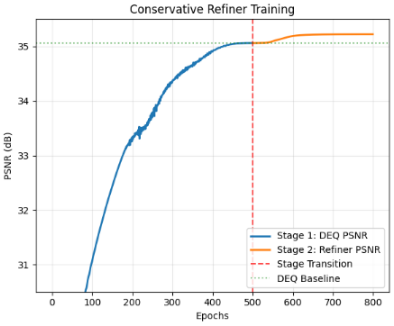

Fig. 4. PSNR (dB) Curve - Astronaut

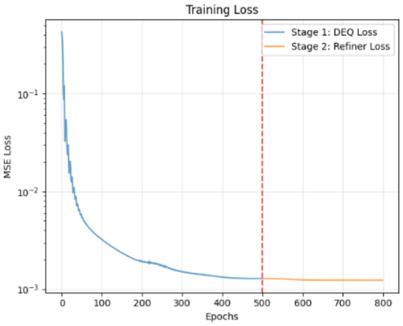

Fig. 5. Loss Curve - Astronaut

One important aspect of DEQs is that they reach a fixed state (converge), and Fig. 4-5 show this around 410 epochs. Even at a fine scale, there is essentially no movement in PSNR or loss past this point. This can potentially be contributed to DEQ's ability to solve globally, where the entire image is reconstructed simultaneously to avoid inconsistencies. However, there is a noticeable rise in PSNR and decrease in loss around 550 epochs, which indicates that the CRR was able to keep the stability of the DEQ with a bump in accuracy. This showcases success by our ***Conservative* Residual Refiner**, in its ability to be stable yet refine the successful DEQ model.

Fig. 6-7 below display the PSNR (dB) and Loss Curves on the image reconstruction of the conjoined DEQ+CRR on the photographer image:

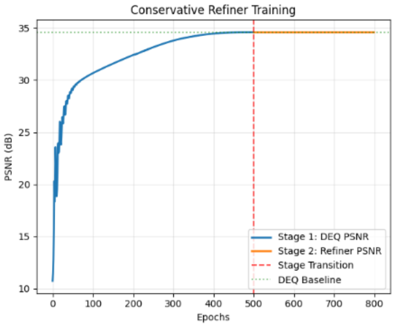

Fig. 6. PSNR (dB) Curve - Photographer

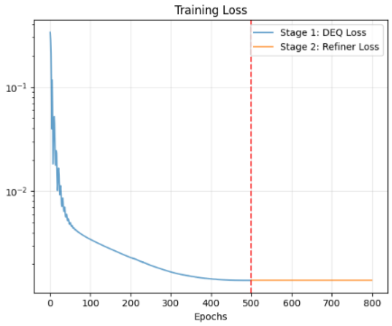

Fig. 7. Loss Curve - Photographer

Both curves stay flat for the photographer's image reconstruction after about 400 epochs. The refiner is deemed unsuccessful because it was unable to raise the reconstruction's PSNR, even lowering it by 0.01. The original astronaut and photographer images, their DEQ reconstruction, and their final reconstruction following the CRR are shown in Fig. 8-9 below:

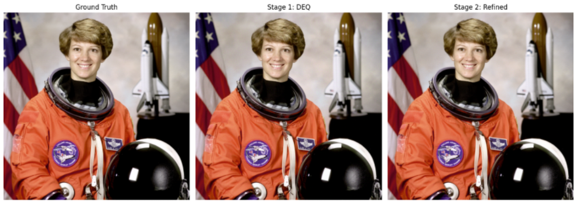

Fig. 8. Astronaut Reconstructions (Original, DEQ, DEQ+CRR)

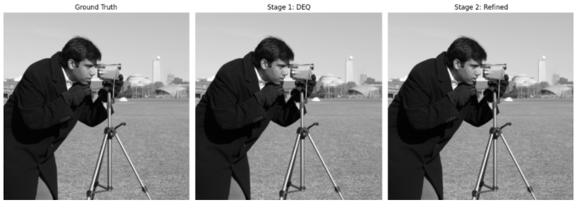

Fig. 9. Photographer Reconstructions (Original, DEQ, DEQ+CRR)

V. Analysis of Residuals and Frequencies

We assessed the spectral complexity and error distribution of the input signals in order to examine the difference in reconstruction fidelity between the two chosen images. We employed three distinct metrics to achieve this: Frequency Spectrum (FS), which is essentially a heatmap displaying the sparsity of the signal, High-Frequency Energy (HFE), which measures the percentage of variance (signal power) outside of the low-frequency DC component, revealing the sparsity of the signal's basis functions, and Significant Residuals (SR), which detects the density of errors that exceed a dynamic 5% range threshold.

A. Spectral and Spatial Data

Each signal had a unique profile according to spectral analysis. The The photographer's image had an HFE of 78.2% and an FS that was primarily purple with highlights of pink and orange. Geometric discontinuities (step-function edges) that require infinite high-frequency components to represent (straightness) are indicated by this broadband energy. A heavy-tailed error distribution at the structural boundaries is indicated by the complexity's 17.0% Significant Residuals. The photographer image's FS and SR are displayed in Fig. 10 below:

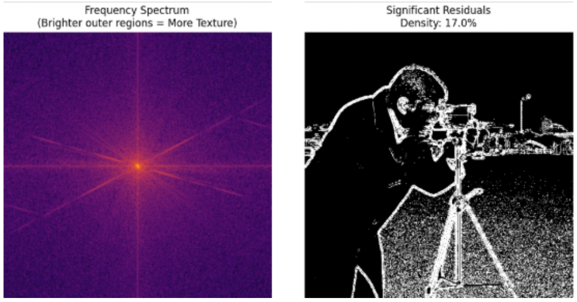

Fig. 10. Frequency Spectrum and Significant Residuals (%) – Photographer

On the other hand, the astronaut registered a lower HFE of 73.3% and a mostly black FS, showing sparsity within the spectrum. This indicates that the signal is dominantly continuous functions (textured). The SR of 0.0% implies that the baseline captured all of the geometry perfectly. Fig. 11 below showcases these metrics:

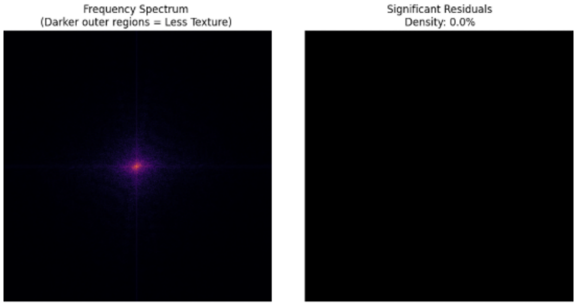

Fig. 11. Frequency Spectrum and Significant Residuals (%) – Astronaut

A complete spectral comparison of the astronaut and photographer is in Fig. 12 below:

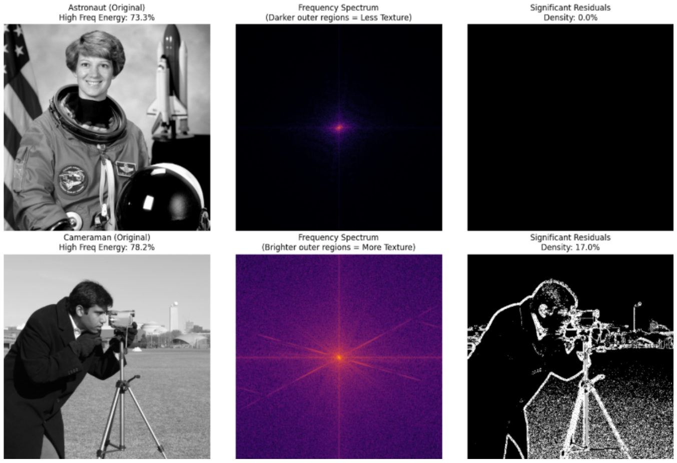

Fig. 12. Spectral Comparison Between Astronaut and Photographer

And described by the following table:

TABLE II
SIGNAL CHARACTERISTICS ANALYSIS

Image Signal	High Freq Energy	Spectrum Visual	Sig. Residuals
Astronaut	73.3%	Mostly black, little purple	0.0%
Photographer	78.2%	Mostly purple, pink/orange	17.0%

B. Hypothesis

We hypothesize that the discontinuity barrier in the photographer image degrades our approach. The following data points explain why.

1) **Broadband Discontinuities (Photographer):** The high HFE and SR confirm that many residuals are geometric edges. Our CRR, a shallow additive network, cannot sinly ”move” an edge. Instead, it can only adjust intensity. Attempting to fit these sharp spectral edges likely resulted in causing the Gibbs Phenomenon, where the refiner introduced some oscillating artifacts (in the form of ringing) instead of a clean edge, leading to a minor decrease in image fidelity (-0.01 dB). Fig. 13 below showcases the Gibbs Phenomenon in general form:

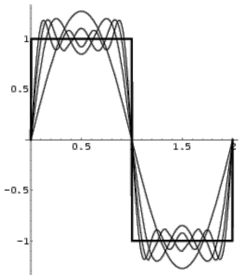

Fig. 13. Gibbs Phenomenon

2) **Continuous Texture (The Astronaut):** The spectral sparsity along with zero SR indicate that the residuals are smooth, continuous gradients. Since neural networks are efficient at approximating continuous functions [1], the refiner successfully modeled these smooth textures without creating oscillating artifacts, which resulted in the observed performance gain (+0.18 dB).

VI. CONCLUSION

In conclusion, our work has focused largely on the usage of Deep Equilibrium Networks for the purpose of image processing tasks. To review, a neural network can be represented by the equation

$$f(X) = \lambda(\alpha X + \beta) \tag{2}$$

and the subtype Deep Equilibrium Neural Network is made possible by a fixed point solving process and can be represented by:

$$z^* = \sigma(Wz^* + u(x) + b) \tag{3}$$

Our three parent papers focused on MRI reconstruction and deblurring, image decompression, and our final paper focuses on structural vs. pixel refinement. All three papers use Deep Equilibrium Networks to complete these tasks, taking advantage of the simulated infinite depth these networks offer. Our unique novelty builds on the work done in these papers to introduce a new model, which aims to build off the shortcomings of prior work. We introduce a two step process, where a traditional DEQ is trained until conversation, and a shallow refiner is separately trained where the DEQ is left off. In this way, we lock in the ability to model simpler data with the “frozen” DEQ, while allowing a refiner to learn from the residuals and build a high-accuracy composite model. Testing was limited to two images, in accordance with the parent paper, however we note that this weakens the generality of our results. Our testing held hyperparameters constant between trails, and used 500 epochs for the baseline DEQ and 300 epochs on the refiner. Our results showed a 0.18 PSNR improvement for the astronaut image and a near-zero difference on the photography image. We believe the continuous texture of the astronaut contributed to improved results, while the discontinuous rigid lines present in the photographer image were optimizable by the refiner, and therefore lead to the same performance as a DEQ without shallow refinement. We believe this work is very beneficial to the scene of DEQ applications in general and specifically image processing and hope it can inspire others to use our research to optimize their own work.

REFERENCES

[1] Z. Geng and J. Z. Kolter, “TorchDEQ: A Library for Deep Equilibrium Models,” arXiv preprint arXiv:2310.18605, 2023.

[2] S. Wang, Y. Teng, and L. Wang, “Deep Equilibrium Object Detection,” arXiv preprint arXiv:2308.09564, 2023.

[3] Y. Zhao, S. Zheng, and X. Yuan, “Deep Equilibrium Models for Video Snapshot Compressive Imaging,” arXiv preprint arXiv:2201.06931, 2023.