Deep Learning Final Project - Reproducing Deep Image Prior

Sean Duffy, Matthew Gregorio, Ikonkar Kaur Khalsa
Khoury College of Computer Sciences
Northeastern University
Boston, MA 02115
Github

December 3, 2024

Abstract

This report explores the *Deep Image Prior* method for image restoration, replicating key results from the original paper. We applied the DIP network to clean, noisy, and shuffled images. DIP restored clean and noisy images effectively, but struggled with shuffled images lacking spatial structure. These findings align with the original paper, confirming the method's effectiveness for structured images and its limitations with disorganized data.

1 Paper to be Reproduced

The paper is titled *Deep Image Prior*. The link to an arxiv copy of the paper is:

`https://arxiv.org/pdf/1711.10925v4`

Authors: Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky

2 Scientific Context of Source Paper

The paper *Deep Image Prior* [1] explores using deep convolutional networks (ConvNets) for image restoration tasks like denoising, super-resolution, and inpainting. It introduces the concept of Deep Image Prior (DIP), where a randomly initialized ConvNet serves as a handcrafted prior, bypassing the need for large-scale pre-training.

DIP leverages the self-similarity and stationarity of natural images. Convolutional filters enforce spatial coherence by exploiting recurring local patterns and structures, enabling the restoration of missing or corrupted image parts.

The restoration process is framed as an optimization problem:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(f_\theta(z), y)$$

Here, f_θ is the deep network model, z is the input (random noise), and $\mathcal{L}$ is the loss function (e.g., MSE) which is described below.

$$E(f_\theta(z), x_0) = \|f_\theta(z) - x_0\|^2$$

Z is a fixed random input tensor sampled from Gaussian distribution, $f_theta(z)$ is output of CNN and x_0 is target image. L2 norm squared loss is used for MSE.

A key insight is that network architecture itself acts as a prior, demonstrating that ConvNets can exploit image structures like self-similarity without relying solely on data-driven learning. Hourglass architectures with skip connections enable the preservation of low- and high-level features, improving restoration performance.

However, the method is computationally expensive, requiring per-image optimization, which limits real-time applicability. Despite this, the paper emphasizes that carefully designed architectures, even with random initialization, can effectively restore images by leveraging their inherent structure without extensive dataset training.

3 Results From Source Paper

The paper demonstrates that neural network architectures alone can act as effective priors for image restoration. Key experiments show that DIP quickly and accurately restores clean, natural images by leveraging self-similarity and stationarity. When applied to noisy images, the network required more iterations but eventually achieved high-quality results. Conversely, for scrambled images or uniform noise with no inherent structure, DIP struggled, producing poor reconstructions due to the lack of spatial coherence to exploit.

To replicate these findings, we tested a similar DIP network on four input types: clean images, noisy images, scrambled images, and noise. For clean images, the network converged quickly, capturing spatial structure and minimizing loss early. With noisy inputs, DIP rejected noise over more iterations, reaching comparable reconstruction quality. For scrambled and noise-only inputs, however, the network failed to produce meaningful outputs, with the loss plateauing early due to the absence of exploitable structure.

These observations confirm that DIP excels with structured inputs like clean and noisy images while failing with disorganized ones. The results reinforce the importance of spatial coherence in the network's ability to reconstruct images effectively.

4 Details of Our Implementation

To replicate the findings from the *Deep Image Prior* paper, we implemented a "U-shaped" neural network model based on the DIP framework and tested its ability to reconstruct various target images.

First, we prepared the data by using a clean image as the starting point. The image was resized to 512×512 pixels and converted into a tensor. From this, we created four target images: a clean image, a noisy image, noise and a shuffled image. The noisy image was generated by adding Gaussian noise, N(0,0.1), to the clean image, while the shuffled image was created by randomly permuting the pixels of the clean image, thereby destroying its spatial structure. The random noise was created using uniform noise, U(0, 1). These images (1) enabled us to evaluate the DIP model's performance on structured (clean and noisy) and unstructured (shuffled and noise) inputs.

The neural network architecture was implemented following the DIP framework, featuring an encoder-decoder structure with skip connections. Specifically, the model consisted of 5 encoding and decoding layers, each with 128 feature maps and kernel sizes of 3 for both upsampling and downsampling operations. Bilinear upsampling was used to maintain spatial resolution in the decoder. We also experimented with an alternative DIP architecture that incorporated additional configurations of skip connections. Limiting skip connections improved performance, specifically in the earlier layers. Our best model featured skip connections in only layers 3 and 4. Both models were initialized with random weights and trained without any pretraining, in accordance with the methodology outlined in the paper.

For training, we used a randomly initialized tensor `z` as the input to the network. During each training iteration, we perturbed `z` with Gaussian noise (standard deviation $\sigma_p = \frac{1}{30}$) to improve the robustness of the reconstruction process. The training objective was to minimize the Mean Squared Error (MSE) between the model's output and the target image. We conducted four separate training experiments, each running for 10,000 epochs: one for reconstructing the clean image, one for the noisy image, one for noise and one for the shuffled image (3). The Adam optimizer with a learning rate of 0.01 was used to update the network parameters. While loss continues to decrease through 10,000 epochs, the quality of the reconstruction suffers as more noise appears in the output. In order to achieve denoising, monitoring of the training process is necessary. Our best reconstruction of the image of the jack-o'-lantern was seen as epoch 5000 2

Finally, we evaluated the model's performance by tracking the MSE loss across epochs and periodically visualizing the reconstructed images. To compare the network's behavior across the four experiments, we generated a loss-vs-iteration plot with a logarithmic x-axis, which clearly highlighted differences in convergence trends. This approach allowed us to observe the network's ability to model structured inputs, such as clean and noisy images, and its limitations when faced with unstructured targets, such as shuffled images and pure noise.

5 Results of Our Implementation

Figure 1: The 4 different images to be reconstructed with DIP architecture.

Figure 2: Reconstruction of natural image with noise through DIP architecture.

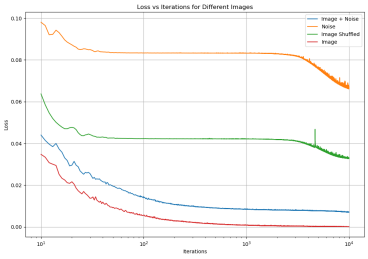

Figure 3: 'Natural' patterns yield faster convergence and less loss outputs than random patterned images.

References

[1] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," *arXiv preprint arXiv:1711.10925*, 2018, Version 4. [Online]. Available: `https://arxiv.org/pdf/1711.10925v4`.