Intermediate Layer Optimization for Inverse Problems using Deep Generative Models

Joseph Dean
University of Texas at Austin
ECE Department
josephdean98@utexas.edu

Giannis Daras
University of Texas at Austin
CS Department
giannisdaras@utexas.edu

Alexandros G. Dimakis
University of Texas at Austin
ECE Department
dimakis@austin.utexas.edu

Abstract

We propose Intermediate Layer Optimization, a novel optimization algorithm for solving inverse problems with deep generative models. Instead of optimizing only over the initial latent code, we progressively change the input layer we optimize over, obtaining successively more expressive generators. We also experiment with different loss functions and utilize a perceptual loss combined with standard mean squared error. We empirically show that our approach outperforms the state-of-the-art inversion methods introduced in StyleGAN2 and PULSE.

1 Introduction

We study how deep generators can be used as priors to solve inverse problems like inpainting, super-resolution and denoising. We focus on unsupervised image reconstruction techniques that rely on a pre-trained generator, building on the general framework introduced by Bora et al. [1] (see also [2] for an overview).

The central optimization problem that appears in unsupervised image reconstruction is the inversion of a deep generative model, i.e. finding a latent code that explains the measurements. This can be performed for different generators, e.g. DCGAN or more recently the powerful StyleGAN2 [3, 4] as shown in the excellent results obtained by PULSE [5]. Unfortunately, inverting a generator with even 4 layers is NP-hard [6].

Prior work [1] used gradient descent to minimize the MSE and showed good empirical performance for numerous inverse problems including inpainting and compressed sensing with random Gaussian measurements using DCGAN. This *does not* work as well for deeper generators e.g. BigGAN as discussed in [7]. PULSE [5] showed significantly better results specifically for super-resolution by improving the latent space optimization and using the powerful generator of StyleGAN2.

We propose a novel optimization method for solving general inverse problems using a technique we call **Intermediate Layer Optimization** (ILO). Our method adaptively changes which layer is optimized, moving from the initial latent code to intermediate layers closer to the pixels. By optimizing intermediate layers we expand the range of the generator to better satisfy the measurements. This has to be done very carefully since intermediate layers can produce non-realistic images, which has to be carefully regularized. We experiment with different loss functions and utilize a perceptual loss combined with standard mean squared error. One benefit of our approach is that training biases seem to be mitigated, as shown in our results.

Background: Inversion is the problem of *projecting an image on the range of a generator*. Given a real image $x \in \mathbb{R}^n$ and a generator $G(z)$, the goal of inversion is to find the latent code $z^* \in \mathbb{R}^k$, so that $G(z^*) \in \mathbb{R}^n$ approximates the given image x as well as possible. The canonical way to perform

NeurIPS 2020 Workshop on Deep Learning and Inverse Problems, virtual.

inversion is to use gradient descent to solve the optimization problem:

$$\min_z \text{dist}\left(G(z), x\right) + R(z), \tag{1}$$

where z is the latent code, x is the given image, $\text{dist}(\cdot,\cdot)$ is a distance function that measures how close the generated image is to the given image and $R(z)$ is a regularization term that forces the latent code to stay close to the original distribution. Several papers have independently introduced this formulation for the inversion [8, 1, 9, 3]. Recent work has generalized this framework to solve more general inverse problems [1, 9, 10, 5, 2]. The general optimization problem is:

$$\min_z \text{dist}\left(f(G(z)), f(x)\right) + R(z), \tag{2}$$

where f is a forward operator that produces measurements from the real data. For example, for inpainting f is a masking operator that hides certain pixels of a given image. MSE has been the most common loss used for solving inverse problems but we show how more careful combinations with perceptual loss functions improves performance.

Using Perceptual Loss: The LPIPS [11] distance metric has been shown to be effective at measuring perceptual differences between two images. This makes it a natural candidate for inversion since the goal is to match the generated image to the observed image. We show in the experiments that incorporating the LPIPS loss function yields superior image reconstructions. The observed image in inpainting is a partially observed image - an image distortion that LPIPS was not trained for. To address this, we minimize the perceptual distance between the generated image and the superimposed reconstruction (see equation 3). In comparison to MSE, we show that a weighted combination of LPIPS and MSE is more effective at image reconstructions under the downsampling and additive noise degradation operator.

Intermediate Layer Optimization: We show how to adapt image inversion to perform image inpainting by searching for latent vectors that match the observed pixels. Our loss function for a given image I is defined as:

$$\min_z \|M \odot I - M \odot G(z)\|_2^2 + \alpha \cdot \text{LPIPS}\left(M \odot I + M^C \odot G(z), G(z)\right) + \beta \cdot R(z), \tag{3}$$

where M is a binary mask, $\odot$ denotes element-wise multiplication, and M^C is the complementary mask, i.e. $M^C(i,j) = \mathbb{1}(M(i,j) = 0)$.

Intermediate Layer Optimization obtains fine-grained image reconstructions. ILO is motivated by the problem of inversion, for which we argue that the optimization scheme of Equation 3 can be further improved. For the problem of inpainting, f is a masking operator that we use to force the model to match only the observable pixels of the given image I. Consider now the case in which the model is observing only the top half of a human face and it is asked to complete it. Solving the optimization problem of Equation 3 usually leads to realistic faces that can not fully match the observed pixels. In other words, super-imposing the original image and the completion leads to observable visual discontinuities. The cause of the problem is that the generator may be unable to produce an image that adequately matches the observations. We therefore need a reconstruction that respects the boundary conditions and matches the observed pixels. ILO is designed to solve exactly this problem by expanding the range of the generator. We gradually optimize consecutive layers of the generator to introduce more flexibility to match observed pixels. Given a deep generator $G = (g_n \circ g_{n-1} \ldots \circ g_2 \circ g_1)(z_1)$, the algorithm runs in rounds, where in each round the initial layer is discarded. At the first round, the algorithm optimizes over the input latent codes. At the second round, we *remove* the first layer from the generator and we optimize over $(g_n \circ g_{n-1} \ldots \circ g_2)(z_2)$, initializing $z_2 = g_1(z_1)$ to stay near the manifold of realistic images before running gradient descent. We repeat this process of removing the initial layer and initializing values from the previous round until the MSE loss becomes very small.

There is a trade-off between the ability to match observed pixels and generation of realistic images. Specifically, if we optimize only over the first latent code layer (as previously done e.g. in [1] and PULSE [5]), we obtain realistic faces that fail to perfectly match the boundary conditions . On the contrary, if we start optimizing only the final layers near the pixels, the generated images do not resemble human faces. However, if we progressively move from the early layers to later layers of the network, we add more flexibility to match the boundary conditions and we are also constrained to move in the manifold of realistic reconstructions. We switch to the next layer when the loss function

flattens, which can be done in an unsupervised manner. As we also show in the Experiments, this simple heuristic works surprisingly well.

Stochastic Noise Addition for Denoising: For denoising, simply inverting a noisy high-resolution image creates grainy reconstructions due to the expressive power of StyleGAN2. We address this issue by minimizing the loss function for a given noisy image I:

$$\min_z \|I - (G(z) + \mathcal{N}(0, \sigma^2))\|_2^2 + \alpha \cdot \text{LPIPS}(I, G(z) + \mathcal{N}(0, \sigma^2)) + \beta \cdot R(z) \tag{4}$$

where $\mathcal{N}(0, \sigma^2)$ is a randomly generated noise added to the generated image at each optimization step. Our method requires knowing the variance of the Gaussian noise added to the original image.

2 Experimental Evaluation

Our experiments are performed using the state of the art StyleGAN2 [4], a ramped-down learning rate scheduler with initial learning rate of 0.1 and the regularization term $R(z)$ defined as the geodesic loss between the latent vectors. We optimize over the 18 different latent vectors and the first 5 noises. We experiment under three different settings: inpainting, denoising, super-resolution. For each task, we use the ideas introduced in the previous section and we compare against the best previously proposed techniques that are based on a formulation of the inversion problem for the image reconstruction task. Namely, for inpainting we use ILO and the formulation of Equation 3, for denoising we use Stochastic Noise Addition and the formulation of Equation 4. For all experiments we use a combination of LPIPS [11] and MSE. We attempt reconstruction of high-resolution (1024×1024) real and generated images.

Figure 1: Inpainting of real images even outside the test set (images from the web), all optimizing the StyleGAN2 generator layers. The LPIPS, MSE+LPIPS columns are an ablation study of the benefits of each innovation. As shown, the combination of ILO with our new loss consistently gives better reconstructions. For all the experiments: $\alpha = 1.5$, $\beta = 0.1$.

Inpainting. We first show results for the problem of inpainting and compare our framework consisting of ILO and the formulation of Equation 3 against the natural extension of PULSE [5] for inpainting. The results are shown in Figure 1. The first four images have masks that were chosen to remove important features and structural integrity of the facial structure. As shown, the combination of ILO and our new loss function consistently gives better completions compared to the baselines. Even when the observations are extremely sparse (e.g. in the last row we observe only 5% of the pixels), our solution produces accurate reconstructions of the original image. To further validate the ILO innovation, we measure the MSE loss between the real and the generated images when we use: (i) just PULSE, (ii) ILO applied for different number of rounds. The results are shown in Figure 2. As shown in the plot, we switch to the next layer in ILO when the MSE error flattens, achieving lower MSE.

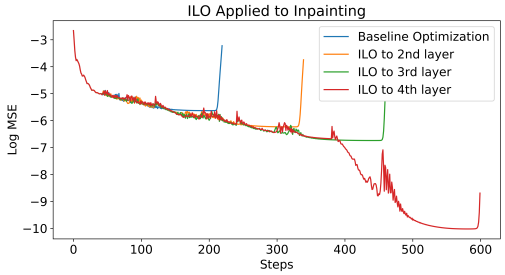

Figure 2: Log MSE achieved through ILO for image reconstruction of masked images.

Denoising. For denoising, we use the Stochastic Addition of Noise framework we introduced earlier together with the ILO method. Previously proposed techniques perform poorly for denoising and hence we also compare with BM3D [12] a standard denoising method. Results are shown in Figure 3 for zero-mean additive Gaussian noise with $\sigma = 25$.

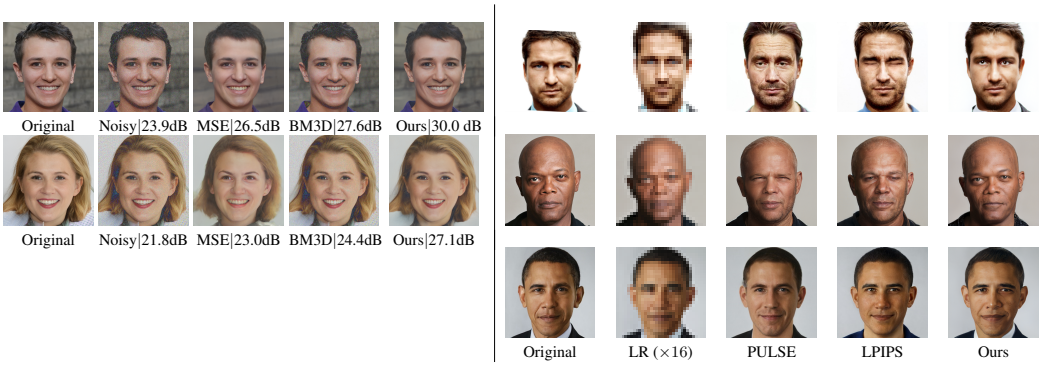

Figure 3: **Left:** Denoising. Gaussian noise ($\sigma = 25$, known) is added to the original image and recovered with various methods. The MSE images indicate the reconstructed images obtained by inverting the noisy image. **Right:** Super-resolution. Many biased reconstructions can be corrected by applying ILO on the weighted combination of MSE and LPIPS.

Super-resolution. Finally, we compare our ILO method for super-resolution with PULSE [5]. Results are shown in Figure 3 (right-part). As shown in the Figure, ILO leads to successful reconstructions. Further, ILO seems to lead to less biased reconstructions since it expands the expressive range of the generator by optimizing intermediate layers.

References

[1] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In *ICML*, 2017.

[2] Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. In *IEEE Journal on Selected Areas in Information Theory (JSAIT)*, 2020.

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[5] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020.

[6] Qi Lei, Ajil Jalal, Inderjit S. Dhillon, and Alexandros G. Dimakis. Inverting deep generative models, one layer at a time. In *NeurIPS*, 2019.

[7] Giannis Daras, Augustus Odena, Han Zhang, and Alexandros G. Dimakis. Your local gan: Designing two dimensional local attention mechanisms for generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[8] Zachary C. Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. In *ICLR*, 2017.

[9] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them all–solving linear inverse problems using deep projection models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5888–5897, 2017.

[10] Maya Kabkab, Pouya Samangouei, and Rama Chellappa. Task-aware compressed sensing with generative adversarial networks. In *AAAI*, 2018.

[11] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[12] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. In *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, volume 6064, page 606414. International Society for Optics and Photonics, 2006.