Premature Termination Strategy for Deep Image Prior

Abstract

Deep Image Prior (DIP) and its variations have demonstrated significant promise in addressing inverse problems in computational imaging, without the need for separate training data. Often, practical DIP models are significantly overparameterized. These models initially capture the intended visual content during the learning phase and subsequently incorporate potential modeling and observational noise, demonstrating a pattern of initial learning followed by overfitting (ELTO). Consequently, the practical application of DIP depends on an early stopping (ES) mechanism capable of identifying this transitional period. Most previous DIP research in computational imaging has focused on demonstrating the models' potential by reporting peak performance against ground truth, without providing practical methods to achieve near-peak performance without access to ground truth. This paper aims to overcome this practical limitation of DIP by introducing an efficient ES strategy that reliably identifies near-peak performance across various computational imaging tasks and DIP variants. This ES method, based on the running variance of intermediate reconstructions in DIP, not only surpasses existing methods that are limited to specific conditions but also maintains its effectiveness when combined with techniques aimed at reducing overfitting.

1 Introduction

Inverse problems (IPs) are widespread in the field of computational imaging, encompassing tasks from fundamental image denoising, super-resolution, and deblurring to complex 3D reconstruction and significant challenges in scientific and medical imaging. Despite the variety of settings, all these problems involve recovering a visual object x from an observation y = f(x), where f represents the forward physical process. Usually, these visual IPs are underdetermined, meaning x cannot be uniquely ascertained from y. This ambiguity is further complicated by potential modeling inaccuracies (such as using a linear f to approximate a nonlinear process) and observational noise (like Gaussian or shot noise), represented as y ˘2248 f(x). To address nonuniqueness and enhance stability against noise, researchers often integrate a range of problem-specific priors on x when formulating IPs.

2 Related Work

There are three primary methods to counteract the overfitting of DIP models. The first one is Regularization: Overfitting is lessened by limiting the size of G˘03b8 to the underparameterization range. Layer-wise weights or the network Jacobian are regularized to regulate the network capacity. The total-variation norm or trained denoisers are used as additional regularizers R(G˘03b8(z)). To prevent overfitting, these techniques need the proper amount of regularization, which varies depending on the kind and degree of noise. They may nevertheless cause overfitting if the regularization level is incorrect. Furthermore, even when they are successful, the performance peak is delayed until the last few iterations, which frequently increases the computing cost by several times. The second method is Noise modeling: In their optimization objective, sparse additive noise is explicitly represented. Regularizers and ES criteria are created especially for Gaussian and shot noise. Subgradient techniques using decreasing step size schedules are being investigated for impulse noise with the ˘21131 loss, and they have shown some early promise. These techniques are ineffective outside of the noise types and levels that they are designed to address, and our understanding of the noise in a particular visual IP is often constrained. The third method is Early stopping (ES): Progress is tracked using a ratio of no-reference blurriness and sharpness, however, as the authors point out, the criterion is only applicable to their modified DIP models. It is unclear how to apply the noise-specific regularizer and ES criterion to unknown noise types and levels. It is suggested to monitor DIP reconstruction by training a coupled autoencoder. Although it performs similarly to ours, the additional autoencoder training significantly increases the overall processing time. By dividing the elements of y into "training" and "validation" sets, it is possible to simulate validation-based ES in supervised learning. However, in IPs, particularly nonlinear ones (such as blind image deblurring (BID), where y ˘2248 k ˘2217 x and ˘2217 denotes linear convolution), elements of y may not be i.i.d., which could impair the effectiveness of validation. Furthermore, withholding a portion of the observation in y can significantly diminish peak performance.

3 Methodology

We advocate for the ES approach because, even when effective, regularization and noise modeling techniques frequently fail to enhance peak performance; instead, they extend it to the final iterations, potentially requiring ten times more iterations than would be necessary to reach the peak in the original DIP models. Furthermore, both approaches necessitate extensive knowledge of the noise type and level, which is often unavailable for most applications. If their essential models and hyperparameters are not appropriately configured, overfitting is likely to persist, and ES will still be necessary. This paper introduces a novel ES criterion applicable to various DIP models, based on monitoring the trend of the running variance in the reconstruction sequence.

Detecting transition by running variance:

Our lightweight method only involves computing the VAR curve and numerically detecting its valleyž2014 the iteration stops once the valley is detected. To obtain the curve, we set a window size parame- ter W and compute the windowed moving variance (WMV). To robustly detect the valley, we introduce a patience number P to tolerate up to P consecutive steps of variance stagnation. Obviously, the cost is dominated by the calculation of variance per step, which is O(W N) (N is the size of the visual object). In comparison, a typical gradient update step for solving Eq. (2) costs at least ž126(|ǒ3b8|N), where |ǒ3b8| is the number of parameters in the DNN Gǒ3b8. Since |ǒ3b8| is typically much larger than W (default: 100), our running VAR and detection incur very little compu- tational overhead.

4 Experiments

ES-WMV is tested for DIP in a variety of linear and nonlinear IPs, including image denoising, inpainting, demosaicing, super-resolution, MRI reconstruction, and blind image deblurring. ES-WMV is also systematically assessed for major DIP variants, such as deep decoder, DIP-TV, and GP-DIP, for image denoising. It is shown to be a dependable helper in identifying effective ES points. The specifics of the DIP variants are covered in Appendix A.5. In addition, ES-WMV is contrasted with the primary rival techniques, such as DF-STE, SV-ES, DOP, SB, and VAL. The specifics of the primary ES-based techniques are found in Appendix A.6. Reconstruction quality is evaluated using both PSNR and SSIM, and detection performance is shown using PSNR and SSIM gaps, which are the differences between our detected and peak values.

4.1 Image Denoising

The majority of earlier research on DIP overfitting has concentrated on image denoising and often assessed their techniques using only one or two forms of noise with modest noise levels, such as low-level Gaussian noise. We use the traditional 9-image dataset for each noise type, and we create two noise levelsž2014low and highž2014for each.

4.2 Image Super-Resolution

In this task, we try to recover a clean im- age x0 from a noisy downsampled ver- sion y = Dt(x0) + ǒ3f5, where Dt(ǒ0b7) : [0, 1]3ǒ0d7tHǒ0d7tW ž192 [0, 1]3ǒ0d7Hǒ0d7W is a down- sampling operator that resizes an im- age by the factor t and ǒ3f5 models ex- tra additive noise. We consider the fol- lowing DIP-reparametrized formulation . = ž225Dt(Gǒ3b8(z)) ž212 yž2252 minǒ3b8 ž113(ǒ3b8) F , where Gǒ3b8 is a trainable DNN parameterized by ǒ3b8 and z is a frozen random seed. Then we conduct experiments for 2ǒ0d7 super- resolution with low-level Gaussian and impulse noise. We test our ES-WMV for DIP and a state-of-the-art zero-shot method based on pre-trained diffusion modelž2014DDNM+ on the standard super-resolution dataset Set14, as shown in Tab. 5, Fig. 11, and Appendix A.7.9. We note that DDNM+ relies on pre-trained models from large external training datasets, while DIP does not. We observe that (1) Our ES-WMV is again able to detect near-peak performance for most images: the average PSNR gap is ž264 1.50 and the average SSIM gap is ž264 0.07; (2) DDNM+ is sensitive to the noise type and level: from Tab. 5, DDNM+ trained assuming Gaussian noise level ǒ3c3y = 0.12 outperforms DIP and DIP+ES-WMV when there is Gaus- sian measurement noise at the level ǒ3c3y = 0.12, which is unrealistic in practice, as the noise level is often unknown beforehand. When the noise level is not set correctly, e.g., as ǒ3c3y = 0 in the DDNM+ (ǒ3c3y = .00) row of Tab. 5, the performance of DDNM+ is much worse than that of DIP and DIP+ES-WMV. Also, for super-resolution with impulse noise, DIP is also a clear winner that leads DDNM+ by a large margin; and (3) in Appendix A.8, we show that DDNM+ may also suffer from the overfitting issue.

4.3 MRI Reconstruction

We also test ES-WMV on MRI reconstruction, a typical linear IP with a nontrivial forward mapping: y ž248 F(x), where F is the subsampled Fourier operator, and we use ž248 to indicate that the noise encountered in practical MRI imaging may be hybrid (e.g., additive, shot) and uncertain. Here, we take the 8-fold undersampling and parameterize x using ž201cConv-Decoderž201d, a variant of deep decoder. Due to the heavy over-parameterization, overfitting occurs and ES is needed.

4.4 Blind Image Deblurring

In BID, a blurry and noisy image is given, and the goal is to recover a sharp and clean image. The blur is mostly caused by motion and/or op- tical non-ideality in the camera, and the forward process is often modeled as y = k ∆2217 x + n, where k is the blur kernel, n models additive sensory noise, and ∆2217 is linear convolution to model the spa- tial uniformity of the blur effect. BID is a very challenging visual IP due to bilin- earity: (k, x) 7∆2192 k ∆2217 x. Recently, researchers have tried to use DIP models to solve BID by modeling k and x as two separate DNNs, i.e., min∆03b8k,∆03b8x ∆2225y ∆2212 G∆03b8k (zk) ∆2217 G∆03b8x(zx)∆22252 2 + ∆03bb∆2225∆2207G∆03b8x (zx)∆22251/∆2225∆2207G∆03b8x (zx)∆22252, where the regular- izer is to promote sparsity in the gradient domain for the reconstruction of x, as stan- dard in BID. We follow previous work and choose a multilayer perceptron (MLP) with softmax activation for G∆03b8k , and the canonical DIP model (CNN-based encoder-decoder architecture) for G∆03b8x(zx). We change their regularizer from the original ∆2225∆2207G∆03b8x (zx)∆22251 to the current, as their original formulation is tested only at a very low noise level ∆03c3 = 10∆22125 and no overfitting is observed. We set the test with a higher noise level ∆03c3 = 10∆22123, and find that its original formulation does not work.

5 Results

Table 1: Summary of performance of our DIP+ES-WMV and competing methods on image denoising and blind image deblurring (BID). ∆2713: working reasonably well (PSNR ∆2265 2dB less of the original DIP peak); -: not working well (PSNR ∆2264 2dB less of the original DIP peak): N/A: not applicable (i.e., we do not perform comparison due to certain reasons). Note that DF-STE, DOP, and SB are based on modified DIP models.

	Image denoising								BID	
	Gaussian		Impulse		Speckle		Shot		Real world	
	Low	High	Low	High	Low	High	Low	High	Low	High
DIP+ES-WMV (Ours)	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713
DIP+NR-IQMs	-	-	-	-	-	-	-	-	N/A	N/A
DIP+SV-ES	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	N/A	N/A
DIP+VAL	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	∆2713	-	-
DF-STE	∆2713	∆2713	N/A	N/A	N/A	N/A	∆2713	∆2713	N/A	N/A
DOP	N/A	N/A	∆2713	∆2713	N/A	N/A	N/A	N/A	N/A	N/A
SB	∆2713	∆2713	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 2: ES-WMV (our method) on real-world image denoising for 1024 images: mean and (std) on the images. (D: detected)

SSIM Gap	∆2113 (loss)	PSNR (D)	PSNR Gap	SSIM (D)
MSE	34.04 (3.68)	0.92 (0.83)	0.92 (0.07)	0.02 (0.04)
∆21131	33.92 (4.34)	0.92 (0.59)	0.93 (0.05)	0.02 (0.02)
Huber	33.72 (3.86)	0.95 (0.73)	0.92 (0.06)	0.02 (0.03)

Table 3: Wall-clock time (secs) of DIP and three ES methods per epoch on NVIDIA Tesla K40 GPU : mean and (std). The total wall clock time should contain both DIP and a certain ES method.

DIP	SV-ES	ES-WMV	ES-EMV
0.448 (0.030)	13.027 (3.872)	0.301 (0.016)	0.003 (0.003)

The results of our experiments are summarized in the tables above. Table 1 shows the performance of our DIP+ES-WMV method against competing methods for image denoising and BID. Table 2 reports the performance of ES-WMV on real-world image denoising for 1024 images. Table 3 compares the wall-clock time of DIP and three ES methods per epoch. Table 4 compares ES-WMV and SB for image denoising on the CBSD68 dataset. Table 5 compares ES-WMV for DIP and DDNM+ for 2∆00d7 image super-resolution. Table 6 shows the performance of ConvDecoder on MRI reconstruction. Table 7 compares BID detection between ES-WMV and VAL on the Levin dataset. Table 8 compares DIP with ES-WMV vs. DOP on impulse noise. Table 9 compares ES-WMV for DIP and DDNM+ for denoising images with medium-level Gaussian and impulse noise. Table 10 compares detection performance between DIP with ES-WMV and DIP with ES-EMV for real image denoising on 1024 images. Table 11 compares detection performance between DIP with ES-WMV and DIP with ES-EMV for real image denoising on the PolyU dataset. Table 12 shows the performance of DIP with ES-WMV for image inpainting.

Table 4: Comparison between ES-WMV and SB for image denoising on the CBSD68 dataset with varying noise level ŏ3c3. The higher PSNR detected and earlier detection are better, which are in red: mean and (std).

	ŏ3c3 = 15		ŏ3c3 = 25		ŏ3c3 = 50	
	PSNR	Epoch	PSNR	Epoch	PSNR	Epoch
WMV	28.7(3.2)	3962(2506)	27.4(2.6)	3068(2150)	24.2(2.3)	1548(1939)
SB	29.0(3.1)	4908(1757)	27.3(2.2)	5099(1776)	23.0(1.0)	5765(1346)

Table 5: Comparison of ES-WMV for DIP and DDNM+ for 2ŏ0d7 image super-resolution with low-level Gaussian and impulse noise: mean and (std). The highest PSNR and SSIM for each task are in red. In particular, we set the best hyperparameter for DDNM+ (ŏ3c3y = 0.12), which is unfair for the DIP + ES-WMV combination as we fix its hyperparameter setting.

	PSNR		SSIM	
	Gaussian	Impulse	Gaussian	Impulse
DIP (peak)	22.88 (1.58)	28.28 (2.73)	0.61 (0.09)	0.88 (0.06)
DIP + ES-WMV	22.11 (1.90)	26.77 (3.76)	0.54 (0.11)	0.86 (0.06)
DDNM+ (ŏ3c3y = .12)	25.37 (2.00)	18.50 (0.68)	0.74 (0.11)	0.50 (0.08)
DDNM+ (ŏ3c3y = .00)	16.91 (0.42)	16.59 (0.34)	0.31 (0.09)	0.49 (0.06)

6 Conclusion

This paper introduces an innovative ES detection approach, ES-WMV, along with its variant, ES-EMV, which has demonstrated robust performance across a range of visual IPs and different DIP variations. In contrast to most competing ES methods that are specific to certain types of noise or DIP models and have limited applicability, our method exhibits broad effectiveness. While there is a method with comparable performance, it significantly increases processing time. Another method, validation-based ES, performs well in simple denoising tasks but falls short in more complex nonlinear IPs like BID.

Table 6: ConvDecoder on MRI reconstruction for 30 cases: mean and (std). (D: Detected)

PSNR(D)	PSNR Gap	SSIM(D)	SSIM Gap
32.63 (2.36)	0.23 (0.32)	0.81 (0.09)	0.01 (0.01)

Table 7: BID detection comparison between ES-WMV and VAL on the Levin dataset for both low-level and high-level noise: mean and (std).Higher PSNR is in red and higher SSIM is in blue. (D: Detected)

	Low Level		High Level	
	PSNR(D)	SSIM(D)	PSNR(D)	SSIM(D)
WMV	28.54(0.61)	0.83(0.04)	26.41(0.67)	0.76(0.04)
VAL	18.87(1.44)	0.50(0.09)	16.69(1.39)	0.44(0.10)

Table 8: DIP with ES-WMV vs. DOP on impulse noise: mean and (std). (D: Detected)

	Low Level		High Level	
	PSNR	SSIM	PSNR	SSIM
DIP-ES	31.64 (5.69)	0.85 (0.18)	24.74 (3.23)	0.67 (0.19)
DOP	32.12 (4.52)	0.92 (0.07)	27.34 (3.78)	0.86 (0.10)

Table 9: Comparison of ES-WMV for DIP and DDNM+ for denoising images with medium-level Gaussian and impulse noise: mean and (std). The highest PSNR and SSIM for each task are in red. In particular, we set the best hyperparameter for DDNM+ (ŏ3c3y = 0.18), which is unfair for the DIP + ES-WMV combination as we fix its hyperparameter setting.

	PSNR		SSIM	
	Gaussian	Impulse	Gaussian	Impulse
DIP (peak)	24.63 (2.06)	37.75 (3.32)	0.68 (0.06)	0.96 (0.10)
DIP + ES-WMV	23.61 (2.67)	36.87 (4.29)	0.60 (0.13)	0.96 (0.10)
DDNM+ (ŏ3c3y = .18)	26.93 (2.25)	22.29 (3.00)	0.78 (0.07)	0.62 (0.12)
DDNM+ (ŏ3c3y = .00)	15.66 (0.39)	15.52 (0.43)	0.25 (0.10)	0.30 (0.10)

Table 10: Detection performance comparison between DIP with ES-WMV and DIP with ES-EMV for real image denoising on 1024 images from the RGB track of NTIRE 2020 Real Image Denoising Challenge: mean and (std). Higher PSNR and SSIM are in red. (D: Detected)

	PSNR(D)-WMV	PSNR(D)-EMV	SSIM(D)-WMV	SSIM(D)-EMV
DIP (MSE)	34.04 (3.68)	34.96 (3.80)	0.92 (0.07)	0.93 (0.07)
DIP (ǎ1131)	33.92 (4.34)	34.83 (4.35)	0.93 (0.05)	0.94 (0.05)
DIP (Huber)	33.72 (3.86)	34.72 (4.04)	0.92 (0.06)	0.93 (0.06)

Table 11: Detection performance comparison between DIP with ES-WMV and DIP with ES-EMV for real image denoising on the PolyU dataset: mean and (std). Higher PSNR and SSIM are in red. (D: Detected)

	PSNR(D)-WMV	PSNR(D)-EMV	SSIM(D)-WMV	SSIM(D)-EMV
DIP (MSE)	36.83 (3.07)	37.32 (3.82)	0.98 (0.02)	0.98 (0.03)
DIP (ǎ1131)	36.20 (2.81)	36.43 (3.22)	0.97 (0.02)	0.97 (0.02)
DIP (Huber)	36.76 (2.96)	37.21 (3.19)	0.98 (0.02)	0.98 (0.02)

Table 12: DIP with ES-WMV for image inpainting: mean and (std). PSNR gaps below 1.00 are colored as red; SSIM gaps below 0.05 are colored as blue. (D: Detected)

	PSNR(D)	PSNR Gap	SSIM(D)	SSIM Gap
Barbara	21.59 (0.03)	0.20 (0.03)	0.67 (0.00)	0.00 (0.00)
Boat	21.91 (0.10)	1.16 (0.18)	0.68 (0.00)	0.03 (0.01)
House	27.95 (0.33)	0.48 (0.10)	0.89 (0.01)	0.01 (0.00)
Lena	24.71 (0.30)	0.37 (0.18)	0.80 (0.00)	0.01 (0.00)
Peppers	25.86 (0.22)	0.23 (0.05)	0.84 (0.01)	0.02 (0.00)
C.man	25.26 (0.09)	0.23 (0.14)	0.82 (0.00)	0.01 (0.00)
Couple	21.40 (0.44)	1.21 (0.53)	0.63 (0.01)	0.04 (0.02)
Finger	20.87 (0.04)	0.24 (0.17)	0.77 (0.00)	0.01 (0.01)
Hill	23.54 (0.08)	0.25 (0.11)	0.70 (0.00)	0.00 (0.00)
Man	22.92 (0.25)	0.46 (0.11)	0.70 (0.01)	0.01 (0.00)
Montage	26.16 (0.33)	0.38 (0.26)	0.86 (0.01)	0.03 (0.01)