A Principled Self-Referenced Early Stopping Approach for Deep Image Prior

Chaoyan Huang[1,2], Cheng-Han Huang[1], Ismail R. Alkhouri[2,3], Rongrong Wang[1,4]
[1]Department of Computational Mathematics, Science, & Engineering, Michigan State University
[2]Department of Electrical Engineering and Computer Science, University of Michigan
[3]Michigan SPARC, Los Alamos National Laboratory
[4]Department of Mathematical Sciences, Michigan State University

Abstract

Recently, Deep Image Prior (DIP) has demonstrated strong capabilities for solving inverse imaging problems (IIPs) by optimizing a randomly initialized convolutional neural network in a training-data-free regime. However, DIP suffers from overfitting to noisy measurements due to network over-parameterization, making early stopping (ES) essential. The most successful ES method tracks fluctuations in the running variance of the network output to detect overfitting. However, in many applications, these fluctuations may appear prematurely, leading to unstable reconstructions. In this paper, we first show that nearly optimal DIP early stopping can be achieved when two independent noisy copies of the degraded image are available. Motivated by this observation, and since obtaining two fully independent copies is infeasible, we propose an overfitting detection framework based on constructing pseudo self-referenced images, resulting in three IIP-specific algorithms. Our approach is further supported by theoretical results on single-reference validation, pseudo-validation estimation, and the impact of shared noise. Across different IIPs, ranging from natural image restoration to medical image reconstruction, and under varying noise levels and noise types, our methods consistently outperform existing DIP early stopping approaches, all without requiring an accurate estimate of the noise level.

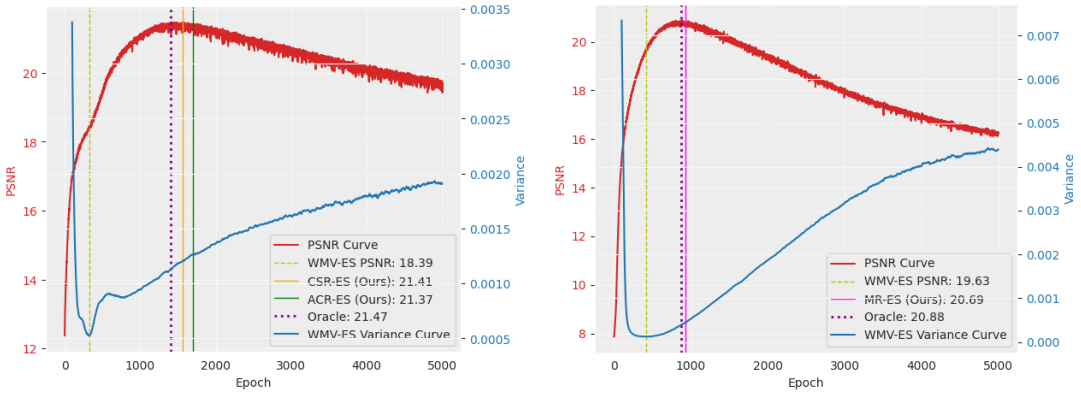
1 Introduction

Inverse Imaging Problems (IIPs) appear in several science and engineering applications, including medical imaging [1, 2], remote sensing [3, 4], and computer vision [5, 6]. IIPs are often ill-posed [7, 8, 9], making exact recovery challenging and requiring advanced reconstruction techniques.

In recent years, several neural network (NN)-based approaches have been developed for solving IIPs [8]. Depending on the availability and type of training data, these methods can be broadly categorized into (*i*) training-data-intensive methods, including supervised [10], self-supervised [11], and generative approaches [12]; and (*ii*) training-free methods such as Implicit Neural Representations (INRs) [13] and the Deep Image Prior (DIP) [14]. Training-data-free methods operate without pre-trained models, making them particularly useful in scenarios where data is scarce, severely degraded, or unlabeled, preventing supervised or generative approaches from providing reliable priors.

INRs provide continuous neural representations of signals and have been successfully applied to several reconstruction tasks [15, 16, 17, 18, 19]. In this paper, however, we focus on DIP methods, which have demonstrated strong performance across a wide range of IIPs [20, 21, 22], in some cases even outperforming training-data-intensive approaches [23, 24].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Natural image (b) CT image

Figure 1: Denoising result on Gaussian noise with noise level 0.26. The oracle is defined as the iteration that achieves the maximum PSNR along the DIP trajectory. The variance-based WMV-ES [25] reaches its global minimum too early and therefore selects a stopping point far from the oracle. In contrast, our CSR-ES, ACR-ES, and MR-ES select iterations much closer to the oracle.

In the original DIP study, the authors in [14] demonstrated that an untrained convolutional neural network (CNN) is able to capture low-frequency image statistics even before optimization. Despite its strong empirical performance [26], DIP remains vulnerable to overfitting due to network over-parameterization, often fitting measurement noise or null-space artifacts of the forward model, as shown for MRI in [27]. Over the past few years, several approaches have been proposed to mitigate this issue which, as surveyed in [26], can be broadly categorized into regularization, network re-parameterization, and early stopping.

Regularization methods, ranging from total variation DIP (TV-DIP) [28] to auto-encoding sequential DIP (aSeqDIP) [24], aim to reduce overfitting. However, these methods can be computationally expensive due to additional optimization terms and variables, and may also be sensitive to regularization hyperparameters.

Early Stopping (ES) [25], which is the focus of this paper, defines stopping criteria that attempt to detect when overfitting begins. Existing DIP ES approaches typically estimate overfitting from either reconstruction trajectory statistics or unbiased risk surrogates. Variance-based methods [25] monitor fluctuations in the DIP output and can be effective in many settings. However, the resulting signal depends not only on overfitting, but also on optimization dynamics and intrinsic image structure. For high-resolution images or images containing substantial high-frequency details, such as textured natural images or medical images with fine anatomical structures, oscillations may appear even before severe overfitting occurs, making the variance curve ambiguous or unstable (see Fig. 1).

Stein's Unbiased Risk Estimator (SURE) [29] can be used for ES, providing a more principled estimate of reconstruction risk. However, SURE-based methods require knowledge of the measurement noise level and are typically more reliable in expectation over multiple realizations. In the per-instance DIP setting, where stopping decisions must be made from a single reconstruction trajectory, these estimates can exhibit high variance and lead to unreliable stopping decisions. These limitations motivate validation-based criteria that directly compare the DIP trajectory against a reference signal not used during fitting. Toward addressing these limitations, we make the following contributions:

- We demonstrate that if two fully independent noisy copies of the degraded measurements are available, then near-optimal overfitting detection becomes achievable. This is supported by theoretical results on single-reference validation, pseudo-validation estimation, and the effect of shared noise.

- Since obtaining *fully* independent degraded measurements is generally infeasible, we propose three task-specific ES algorithms based on pseudo-reference construction: channel-similarity reference ES (CSR-ES) for RGB images, mask-reference ES (MR-ES) for grayscale images such as MRI and CT, and augmented-channel reference ES (ACR-ES) for handling different noise types, including Poisson noise.

- Extensive evaluation across natural image restoration and medical image reconstruction tasks, under varying noise levels and noise types, demonstrates that the proposed methods more reliably detect overfitting compared to existing variance-based and SURE-based DIP ES methods, all without requiring an accurate estimate of the noise level.

1.1 Preliminaries

DIP setup and notation. We consider a noisy measurement $y = A(x) + \eta$, where x is the unknown clean image and A is the forward operator. DIP reconstructs x by optimizing an untrained network $f_\theta(z)$ to fit the measurement,

$$\theta_t \approx \arg\min_\theta \mathcal{L}(A(f_\theta(z)), y),$$

and we denote the resulting reconstruction at iteration t by $\widehat{x}_t(y) = f_{\theta_t}(z)$. Since continued optimization eventually fits measurement noise or null-space artifacts, early stopping chooses a time $\widehat{t}$ and returns $\widehat{x}_{\widehat{t}}(y)$.

2 Reference-Based Early Stopping

DIP early stopping can be viewed as a validation problem. Let $y_1 = x + \eta_1$ be the noisy image used for fitting, and suppose for the moment that an independent noisy reference $y_2 = x + \eta_2$ were available. If $\{\hat{x}_t(y_1)\}_{t=1}^T$ denotes the DIP trajectory trained only on y_1, then

$$\mathbb{E}_{\eta_2}\left[\frac{1}{n}\|\hat{x}_t(y_1) - y_2\|^2 \,\middle|\, y_1\right] = \frac{1}{n}\|\hat{x}_t(y_1) - x\|^2 + \sigma^2.$$

Thus, the reference loss has the same minimizer as the clean-image MSE, up to an additive constant. Moreover, because the loss averages over all pixels, a single independent reference is already sufficient to obtain a concentrated validation signal.

Proposition 2.1 (Single-reference validation, informal)**.** *Assume that the reference noise is independent, zero-mean, and sub-Gaussian. Then, with probability at least* $1 - \delta$*, uniformly over* $t \le T$*,*

$$\frac{1}{n}\|\hat{x}_t(y_1) - y_2\|^2 = \frac{1}{n}\|\hat{x}_t(y_1) - x\|^2 + \sigma^2 + O\left(\sqrt{\frac{\log(T/\delta)}{n}}\right).$$

The formal statement, constants, proof, and the corresponding oracle inequality are given in Appendix A. In standard DIP, however, only one noisy observation is available. We therefore construct a pseudo-reference $\tilde{y}$ from the observed image y. Since $\tilde{y}$ is not fully independent of the image used for fitting, we decompose $y = x + s + \eta, \tilde{y} = x + s + \tilde{\eta}$, where s denotes the component shared by fitting and validation, while η and $\tilde{\eta}$ denote non-shared components. For the pseudo-validation loss

$$\tilde{V}_t := \frac{1}{n}\|\hat{x}_t(y) - \tilde{y}\|^2,$$

we have

$$\mathbb{E}_{\tilde{\eta}}\left[\tilde{V}_t \,\middle|\, y, s\right] = \frac{1}{n}\|\hat{x}_t(y) - (x + s)\|^2 + \sigma_{\tilde{\eta}}^2.$$

Thus pseudo-validation is exact for the effective target $x + s$. Its usefulness for clean-image early stopping depends on whether the shared component s is small, harmless, or not preferentially fitted before the remaining noise.

This gives the main design principle of the paper: a pseudo-reference need not be perfectly independent, but it should reduce the overlap between the component fitted by DIP and the component used for validation. In the noise-fitting model

$$\hat{x}_t(y) \approx x + \alpha_t(s + \eta), \qquad \alpha_t \in [0, 1],$$

which is consistent with DIP's spectral bias when s and η have similar statistics, the expected pseudo-validation loss is minimized near

$$\alpha^\star = \frac{\|s\|^2/n}{\|s\|^2/n + \sigma_\eta^2}.$$

If the construction makes the shared energy a $1/r$ fraction of the non-shared variance, the excess MSE relative to the oracle is $O(1/r^2)$, quadratically tighter than the worst-case $O(1/r)$ bound implied by tracking the effective target $x + s$ above. Modest decorrelation of the pseudo-reference therefore suffices for near-oracle stopping. The full derivation of the noise-fitting model is in Appendix B. The algorithms in Section 3 instantiate this principle using channel separation, held-out masks, and auxiliary corrupted channels.

Scope of the theoretical analysis. The independent-reference result applies to a fitting trajectory that does not use the reference noise. For a dependent pseudo-reference, the risk identity in Proposition B.1 instead concerns the effective target $x + s$. The sharper shared-noise analysis uses a stylized model in which the shared and non-shared components are fitted with a common coefficient. Comparable fitting dynamics are an assumption of this model, not a general property established for DIP optimization. These results motivate our reference constructions, but do not constitute an end-to-end oracle guarantee for all three practical algorithms. In particular, the spectral score used by ACR-ES is a trajectory diagnostic, rather than an unbiased estimator of clean-image risk.

3 Pseudo-Reference Algorithms

Section 2 suggests that ES can be guided by a target that is not fully independent, but whose validation component is sufficiently decorrelated from the component fitted by DIP. We instantiate this principle through three pseudo-reference constructions, chosen according to the available image structure: cross-channel references for RGB images, masks for grayscale images, and auxiliary corrupted images for more challenging noise models. Full algorithmic pseudocodes are given in Appendix C.

3.1 Channel-similarity reference for RGB images

For RGB images, different color channels often share the same image structure while containing partially decorrelated noise. CSR-ES trains DIP on the original RGB image using the standard fitting loss, and uses cross-channel consistency as a post-hoc stopping criterion. Writing

$$y_c = x_c + N_c, \quad c \in \{R, G, B\}, \quad x_c \text{ the clean channel and } N_c \text{ the channel noise},$$

we select the closest channel pair from the noisy image,

$$(i^\star, j^\star) = \arg\min_{i \neq j} \frac{1}{n} \|y_i - y_j\|^2,$$

and use the validation curve

$$V_t^{\mathrm{CSR}} = \frac{1}{n} \|\hat{x}_{t,i^\star} - y_{j^\star}\|^2, \qquad \hat{t}_{\mathrm{CSR}} = \arg\min_t V_t^{\mathrm{CSR}}.$$

The intuition is that selecting a structurally similar reference channel reduces clean cross-channel bias, while the channel-specific noise remains less correlated with the fitted channel.

3.2 Mask-reference early stopping for grayscale images

CSR-ES requires multiple channels and is therefore not directly applicable to grayscale images. MR-ES instead constructs a pseudo-reference by holding out a small subset of pixels. Let $M \in \{0, 1\}^n$ be a high-retention mask, for example $\mathbb{P}(M_i = 1) = 0.98$, and let $H = 1 - M$ be the held-out mask. DIP is trained with the masked loss $\|M \odot (f_\theta(z) - y)\|^2$, so the held-out pixels are excluded from the training loss rather than treated as zero-valued targets. The stopping curve is then

$$V_t^{\mathrm{MR}} = \frac{1}{\sum_i H_i} \|H \odot (\hat{x}_t - y)\|^2, \qquad \hat{t}_{\mathrm{MR}} = \arg\min_t V_t^{\mathrm{MR}}.$$

This construction trades a small loss of training information for a validation signal that is not directly fitted by the network. The held-out validation theorem, its transfer condition to full-image risk, and the implementation requirement are given in Appendix E.

3.3 Augmented-channel reference for flexible pseudo-validation

Mask references can be less reliable when a representative mask is difficult to choose, for example under signal-dependent noise. ACR-ES instead augments the output space of the DIP network. Given the noisy observation y, generate two pseudo-copies

$$y^{(1)} = y + \zeta_1, \qquad y^{(2)} = y + \zeta_2,$$

where ζ_1 and ζ_2 are random noise, and train a $2C$-channel DIP network to fit $Y = \mathrm{concat}(y, y^{(1)})$. The original output channels provide the reconstruction, while the auxiliary channels are monitored against the independent pseudo-copy $y^{(2)}$, whose added perturbation ζ_2 is independent of the perturbation ζ_1 used to construct $y^{(1)}$.

In practice we use a frequency-bias score rather than raw MSE on the auxiliary channel:

$$B_t^{\mathrm{ACR}} = \frac{\sum_\omega \|\omega\|^2 \left|\mathcal{F}(\hat{X}_{t,\mathrm{aux}} - y^{(2)})(\omega)\right|^2}{\sum_\omega \left|\mathcal{F}(\hat{X}_{t,\mathrm{aux}} - y^{(2)})(\omega)\right|^2}, \quad \hat{t}_{\mathrm{ACR}} = \arg\max_{t \geq t_0} B_t^{\mathrm{ACR}} \text{ after a brief burn-in } t_0.$$

This score measures the second moment of the residual spectrum and is more sensitive to the frequency shift that appears when DIP begins fitting noise. For the spectral-bias variant of ACR-ES, we select the stopping point by the maximum of the average-frequency score rather than its minimum. This choice is motivated by the typical spectral evolution of DIP training. The detailed discussion is given in Appendix C. Additional discussion of the augmented-channel construction, the role of τ, and the score choice is given in Appendix F.

3.4 Practical selection and scope

Based on the respective characteristics of these three algorithms, we provide the following guidelines:

(i) CSR-ES is preferred when color or multi-channel observations are available, since it is post hoc and does not alter the standard DIP optimization trajectory.

(ii) MR-ES is most suitable for single-channel observations when a small representative subset of pixels can be withheld without noticeably affecting reconstruction quality, as in medical images where the region of interest is concentrated in the center.

(iii) ACR-ES is the safer default for single-channel observations when the noise type is unknown or known to be Poisson.

When MR-ES is considered but the noise type is uncertain, it can be repeated with several random masks. Low variability in the selected stopping iteration suggests that the mask-based pseudo-reference is reliable, whereas high variability favors ACR-ES.

4 Experimental Results

We evaluate whether the proposed reference-based early-stopping criteria can identify stopping points close to the oracle iteration of DIP optimization. Our experiments are conducted on both natural and medical images. For natural images, we use Set3, Set14, Set18, CBSD68, and Urban100[1]. For medical images, we use MRI15, CT10, PD20, and T2-20from the BrainWeb database[2]. We consider three representative corruption models: Gaussian with noise level 0.26, Poisson with noise level 10, and impulse with noise level 0.1. For Poisson and impulse corruptions, we use their zero-mean versions so that the perturbations do not introduce an additional intensity bias.

The baseline methods include WMV-ES [25] and SURE-based early stopping. For WMV-ES, we follow the setting in [25] and select the stopping iteration according to the local minimum of the windowed variance curve. Specifically, for a DIP trajectory of $5{,}000$ iterations, we use a window size of $1{,}000$ iterations to compute the variance statistic. For SURE, we provide the true noise level used to generate the corrupted observation. This gives SURE a favorable setting, since the noise variance is assumed to be known exactly.

[1] `https://github.com/cszn/IRCNN/tree/master/testsets`

[2] `https://brainweb.bic.mni.mcgill.ca/brainweb/`

Table 1: Average early-stopping performance on natural image datasets under different noise types. All stopping criteria are evaluated on the same ACR trajectory, so the oracle PSNR is shared across WMV-ES, CSR-ES, and ACR-ES. PSNR is reported as mean ± standard deviation, and PSNR Gap denotes the difference from the shared oracle. **This table isolates the stopping rule rather than comparing the original full pipelines. For the latter, please refer to Table 4**

Noise	Dataset	WMV-ES		CSR-ES		ACR-ES		Oracle
		PSNR ↑	Gap ↓	PSNR ↑	Gap ↓	PSNR ↑	Gap ↓	PSNR ↑
Gaussian	Set14	21.86(±2.18)	0.64(±0.56)	**22.27(±1.71)**	0.24(±0.21)	22.26(±1.71)	0.25(±0.13)	22.50(±1.79)
	Set18	21.37(±2.57)	1.27(±1.85)	22.38(±1.59)	0.26(±0.16)	**22.41(±1.58)**	0.23(±0.12)	22.64(±1.65)
	CBSD68	21.40(±2.79)	1.24(±1.04)	**22.48(±2.17)**	**0.16(±0.12)**	22.35(±1.97)	0.30(±0.28)	22.64(±2.18)
	Urban100	18.82(±2.99)	2.75(±2.40)	**21.38(±1.33)**	**0.18(±0.14)**	21.36(±1.31)	0.20(±0.10)	21.56(±1.35)
Poisson	Set14	23.81(±2.93)	0.83(±0.93)	24.30(±2.12)	0.33(±0.41)	**24.52(±2.35)**	**0.11(±0.05)**	24.63(±2.34)
	Set18	25.62(±3.41)	1.08(±1.95)	25.99(±2.02)	0.71(±1.33)	**26.55(±2.33)**	**0.15(±0.10)**	26.70(±2.32)
	CBSD68	23.06(±3.33)	1.74(±1.79)	24.59(±2.29)	0.20(±0.15)	**24.66(±2.33)**	**0.14(±0.09)**	24.80(±2.33)
	Urban100	21.32(±4.00)	2.94(±3.40)	24.07(±1.62)	0.19(±0.13)	**24.08(±1.66)**	**0.18(±0.12)**	24.26(±1.67)
Impulse	Set14	23.33(±2.53)	0.82(±0.83)	**23.77(±1.79)**	**0.22(±0.16)**	23.65(±1.74)	0.49(±0.23)	24.14(±1.88)
	Set18	23.30(±2.64)	1.31(±2.13)	**24.20(±1.55)**	**0.51(±0.34)**	24.09(±1.52)	0.53(±0.21)	24.61(±1.62)
	CBSD68	22.67(±3.01)	1.39(±1.29)	**23.84(±2.17)**	**0.23(±0.19)**	23.63(±2.02)	0.44(±0.25)	24.07(±2.19)
	Urban100	20.05(±3.68)	3.15(±3.13)	**22.96(±1.46)**	**0.24(±0.19)**	22.86(±1.42)	0.34(±0.17)	23.21(±1.51)

We report two quantities: the PSNR at the selected stopping iteration and the corresponding PSNR Gap. The PSNR Gap is defined as $\mathrm{Gap} = \mathrm{PSNR}_{\mathrm{oracle}} - \mathrm{PSNR}_{\mathrm{ES}}$, where $\mathrm{PSNR}_{\mathrm{ES}}$ is the PSNR obtained at the iteration selected by an early-stopping criterion, and $\mathrm{PSNR}_{\mathrm{oracle}}$ is the maximum PSNR along the corresponding DIP trajectory. A smaller gap indicates that the stopping criterion selects an iteration closer to the oracle stopping point. Unless otherwise specified, results are reported as mean ± standard deviation across images.

Our experiments are organized according to the applicability of the proposed criteria. CSR-ES is a post-hoc criterion based on multi-channel consistency, and is therefore most naturally suited for color images. MR-ES constructs a mask-based reference signal, making the reference-based principle applicable to single-channel grayscale images. ACR-ES uses auxiliary corrupted references and provides a more flexible stopping signal for challenging noise models, especially when the reference signal is less reliable. Accordingly, we first evaluate CSR-ES and ACR-ES on natural color images, then evaluate MR-ES on grayscale medical images, then analyze method-specific oracle behavior, and finally test whether the proposed stopping principle generalizes beyond denoising.

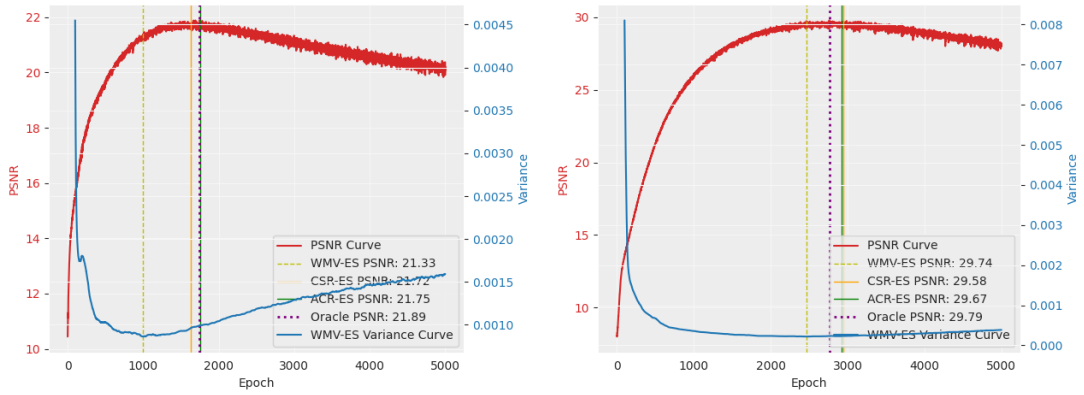
For stopping criteria evaluated on the same DIP trajectory, such as WMV-ES and CSR-ES, the oracle PSNR is shared. For criteria that modify the optimization trajectory, such as MR-ES and ACR-ES, we additionally report method-specific oracle values. This distinction is important because different trajectories may have different attainable oracle PSNR values.

Table 4 compares the full practical pipelines: baseline ES methods on the original DIP trajectory versus our ES method on the modified trajectory. Although the proposed modification may slightly reduce the oracle peak PSNR, it greatly reduces the gap between the selected stopping point and the oracle best point. The reduction in stopping gap compensates for the small loss in peak PSNR, leading to a higher final reconstruction PSNR.

Due to space limitations, the main paper focuses on four representative studies. Additional results in Appendix G provide per-image oracle comparisons, all RGB channel-reference pairs used by CSR-ES, ablations of MSE-based and bias-based reference scores, sensitivity to auxiliary noise levels, early stopping beyond vanilla DIP, and additional qualitative examples.

4.1 Early Stopping on Natural Color Images

We first evaluate the proposed criteria on natural color images. Table 1 is not intended as a full-pipeline comparison. Instead, it compares the stopping methods by evaluating WMV-ES, CSR-ES, and ACR-ES on the same ACR-induced trajectory. Likewise, Table 2 compares WMV-ES and MR-ES by evaluating them on the same MR-induced trajectory. Full practical pipeline comparisons, where each method is evaluated on its own natural trajectory, are reported in Table 4. This setting isolates the effect of the stopping criterion itself, rather than differences caused by different training trajectories.



(a) Gaussian iteration: CSR-ES (1625), ACR-ES (1759), Oracle (1748) (b) Poisson iteration: CSR-ES (2940), ACR-ES (2926), Oracle (2761)

Figure 2: Color image denoising results. Image corrupted by Gaussian noise with noise level 0.26; Poisson noise with noise level 10. Optimizer: Adam with LR 10^{-4}. The number in parentheses after each method indicates the selected iteration. The variance-based criterion often reaches its global minimum too early, leading to premature stopping.

Table 1 reports the average early-stopping performance on Set14, Set18, CBSD68, and Urban100 under Gaussian, Poisson, and impulse noise. Compared with WMV-ES, both CSR-ES and ACR-ES substantially reduce the gap to the shared oracle stopping point across datasets and noise types. The improvement is especially pronounced on structured images, such as Urban100, where WMV-ES often stops too early and leaves a large gap to the oracle. In contrast, the proposed criteria select iterations much closer to the oracle region. CSR-ES is particularly effective under Gaussian and impulse noise, while ACR-ES is more favorable under several Poisson settings, suggesting that auxiliary corrupted references provide a useful stopping signal under signal-dependent corruptions.

Figure 2 provides a trajectory-level view of this behavior. The variance-based WMV score often reaches its minimum before the PSNR peak, leading to premature stopping. By contrast, CSR-ES and ACR-ES select iterations near the oracle region, showing better alignment with the actual overfitting dynamics of DIP.

Appendix G.1 further analyzes all RGB channel-reference pairs used by CSR-ES. The automatically selected pair, computed from the noisy observation only, yields near-oracle stopping across Gaussian, Poisson, and impulse noise. This supports the use of channel similarity as a pseudo-validation signal without access to the clean image. Additional results are provided in Appendix G.2.

4.2 Early Stopping on Grayscale Medical Images

CSR-ES relies on RGB channel consistency and is therefore not directly applicable to single-channel grayscale images. To test whether the proposed reference-based principle can go beyond color images, we first evaluate MR-ES on four grayscale medical datasets. MR-ES constructs a small held-out mask as a reference signal, allowing early stopping without requiring multi-channel information.

Table 2 shows that MR-ES substantially reduces the PSNR Gap under Gaussian and impulse noise across all medical datasets. For example, on the more challenging PD and T2 images, WMV-ES can stop far from the oracle, whereas MR-ES selects iterations much closer to the best attainable point along the trajectory. This demonstrates that the proposed reference-based framework is not limited to RGB images and can be adapted to grayscale medical reconstruction tasks.

However, MR-ES can be less stable under some Poisson settings. This is expected because Poisson noise is signal-dependent, making a simple held-out mask reference less reliable. To address this limitation, we further evaluate ACR-ES, which uses auxiliary corrupted references to construct a more flexible stopping signal. Table 3 reports ACR-ES results on CT10 and T2-20 under Gaussian, Poisson, and impulse noise. Across all settings, ACR-ES substantially improves over WMV-ES and selects stopping points much closer to the oracle. The improvement is especially pronounced under Poisson noise: on CT10, ACR-ES reduces the average gap from 4.46 dB to 0.80 dB, and on T2-20, from 6.41 dB to 1.42 dB. Under Gaussian and impulse noise, ACR-ES also consistently reduces

Table 2: Average early-stopping performance on medical datasets under different noise types. PSNR is reported as mean ± standard deviation across images, and PSNR Gap denotes the difference from the oracle best stopping point.

Noise	Dataset	WMV-ES		MR-ES		Oracle
		PSNR ↑	Gap ↓	PSNR ↑	Gap ↓	PSNR ↑
Gaussian	MRI15	21.41(±0.54)	0.60(±0.38)	**21.60(±0.69)**	**0.41(±0.29)**	22.01(±0.58)
	CT10	19.63(±0.60)	0.94(±0.34)	**20.22(±0.47)**	**0.35(±0.28)**	20.57(±0.43)
	PD20	18.12(±0.92)	2.15(±0.86)	**19.94(±0.36)**	**0.33(±0.24)**	20.27(±0.15)
	T2-20	17.11(±0.55)	2.67(±0.51)	**19.62(±0.19)**	**0.16(±0.09)**	19.78(±0.12)
Poisson	MRI15	**25.37(±0.88)**	**0.76(±0.34)**	24.78(±1.27)	1.35(±1.04)	26.13(±0.71)
	CT10	23.45(±0.10)	3.15(±1.09)	**24.56(±1.30)**	**2.03(±0.85)**	26.60(±1.14)
	PD20	**22.71(±2.35)**	**2.93(±2.41)**	21.15(±0.37)	4.49(±0.37)	25.63(±0.43)
	T2-20	18.75(±1.22)	7.56(±1.40)	**23.26(±0.47)**	**3.05(±0.36)**	26.32(±0.55)
Impulse	MRI15	23.29(±0.76)	0.65(±0.48)	**23.49(±0.52)**	**0.46(±0.47)**	23.94(±0.45)
	CT10	21.55(±0.81)	1.58(±0.36)	**22.82(±0.70)**	**0.31(±0.17)**	23.13(±0.68)
	PD20	21.55(±1.78)	1.83(±1.81)	**23.12(±0.28)**	**0.26(±0.19)**	23.38(±0.21)
	T2-20	17.61(±0.72)	4.59(±0.78)	**21.89(±0.26)**	**0.31(±0.16)**	22.20(±0.21)

Table 3: Average early-stopping performance of ACR-ES on CT10 and T2-20 under different noise types. PSNR is reported as mean ± standard deviation across images, and PSNR Gap denotes the difference from the oracle best stopping point.

Noise	Dataset	WMV-ES		ACR-ES		Oracle
		PSNR ↑	Gap ↓	PSNR ↑	Gap ↓	PSNR ↑
Gaussian	CT10	19.61(±0.68)	1.00(±0.43)	**20.38(±0.51)**	**0.23(±0.09)**	20.61(±0.48)
	T2-20	16.61(±0.72)	3.11(±0.71)	**19.50(±0.23)**	**0.21(±0.15)**	19.71(±0.16)
Poisson	CT10	23.81(±1.24)	4.46(±1.51)	**27.47(±1.28)**	**0.80(±0.54)**	28.27(±1.58)
	T2-20	21.13(±3.73)	6.41(±3.62)	**26.12(±0.80)**	**1.42(±0.64)**	27.54(±0.67)
Impulse	CT10	20.82(±1.00)	2.07(±0.47)	**22.28(±0.85)**	**0.60(±0.37)**	22.88(±0.69)
	T2-20	17.35(±1.30)	4.62(±1.19)	**21.39(±0.34)**	**0.59(±0.37)**	21.98(±0.29)

the gap, showing that auxiliary corrupted references provide a reliable stopping signal for grayscale medical images.

Appendix G.3 provides full per-image method-specific oracle comparisons, where the same trends hold across representative CT and T2 images.

4.3 Method-specific Oracle Analysis

For criteria that modify the reference construction or induce a different optimization trajectory, a small PSNR Gap alone is not sufficient. The gap could become small simply because the attainable oracle PSNR of that trajectory is degraded. We therefore perform a method-specific oracle analysis, where each method is evaluated against the oracle PSNR of its own trajectory.

Table 4 summarizes the average ES PSNR, method-specific oracle PSNR, and oracle gap on representative natural and medical image groups. Each PSNR entry is reported as ES PSNR (Oracle PSNR), so the selected reconstruction quality and the best attainable quality along the corresponding trajectory can be compared simultaneously. On natural images, the proposed CSR-ES and ACR-ES achieve much smaller gaps than SURE-ES and WMV-ES while maintaining oracle PSNR values comparable to the competing trajectories. For example, under Gaussian noise, CSR-ES reduces the average gap on randomly selected images from Set3 and Set14 to 0.17 dB, compared with 1.16

Table 4: Average PSNR and oracle gap across image groups. For each method, the value in parentheses is the average peak PSNR. The $\pm$ values denote standard deviations across images.

Noise	Image group	SURE-ES		WMV-ES		MR-ES		CSR-ES		ACR-ES	
		ES PSNR (Oracle) ↑	Gap ↓	ES PSNR (Oracle) ↑	Gap ↓	ES PSNR (Oracle) ↑	Gap ↓	ES PSNR (Oracle) ↑	Gap ↓	ES PSNR (Oracle) ↑	Gap ↓
Gaussian	Set3/Set14	21.09(22.24)±1.43	1.16±1.05	20.84(21.91)±1.86	1.07±0.83	21.46(21.79)±1.26	0.33±0.19	**21.62(21.78)±1.39**	**0.17±0.08**	21.61(21.78)±1.29	0.18±0.11
	CT10/T2-20	17.37(20.39)±2.03	3.02±2.08	18.39(20.31)±1.90	1.92±1.37	19.81(20.13)±0.41	0.32±0.29	–		**20.01(20.18)±0.57**	**0.17±0.05**
Poisson	Set3/Set14	22.30(23.31)±2.34	1.00±0.93	22.06(22.86)±2.42	0.80±0.82	22.62(22.86)±1.92	0.24±0.15	23.52(23.66)±1.99	0.14±0.08	**23.53(23.66)±2.03**	**0.12±0.05**
	CT10/T2-20	21.68(26.02)±2.40	4.34±2.47	21.00(26.39)±2.76	5.38±2.62	23.77(26.01)±0.77	2.24±0.88	–		**26.55(27.39)±1.17**	**0.84±0.42**
Impulse	Set3/Set14	20.35(23.91)±2.12	3.56±2.55	22.45(23.42)±2.39	0.97±0.97	22.94(23.26)±1.64	0.32±0.17	**23.16(23.31)±1.68**	**0.15±0.09**	22.89(23.31)±1.50	0.42±0.28
	CT10/T2-20	14.59(22.83)±0.47	8.24±0.44	20.68(22.83)±1.86	2.16±1.29	**22.27(22.60)±0.70**	**0.33±0.17**	–		21.92(22.35)±0.82	0.42±0.40

Table 5: Extension to super-resolution and inpainting on Urban10. WMV-ES and CSR-ES are evaluated on the same standard DIP trajectory; therefore, they share the same oracle PSNR. PSNR and Gap are averaged over 10 images, where Gap denotes Oracle PSNR minus ES PSNR.

Task	Setting	Noise	WMV-ES		CSR-ES		Oracle
			PSNR ↑	Gap ↓	PSNR ↑	Gap ↓	PSNR ↑
SR	×4	Gaussian	16.10(±2.79)	1.20(±1.59)	**17.24(±1.32)**	**0.07(±0.08)**	17.31(±1.38)
		Poisson	16.29(±2.85)	1.60(±1.70)	**17.84(±1.25)**	**0.05(±0.05)**	17.89(±1.29)
		Impulse	16.47(±3.05)	1.69(±1.84)	**18.11(±1.37)**	**0.05(±0.04)**	18.16(±1.38)
	×8	Gaussian	15.29(±2.45)	0.36(±0.39)	**15.60(±2.11)**	**0.05(±0.05)**	15.65(±2.15)
		Poisson	15.71(±2.60)	0.39(±0.53)	**16.03(±2.15)**	**0.07(±0.06)**	16.09(±2.17)
		Impulse	15.81(±2.67)	0.45(±0.61)	**16.17(±2.18)**	**0.09(±0.08)**	16.26(±2.21)
Inpainting	50%	Gaussian	17.07(±3.41)	3.63(±3.61)	**19.53(±0.70)**	**1.16(±0.92)**	20.69(±1.03)
		Poisson	17.29(±3.52)	4.30(±4.00)	**21.16(±0.91)**	**0.43(±0.22)**	21.59(±0.92)
		Impulse	17.44(±3.79)	4.61(±4.11)	**21.11(±1.16)**	**0.94(±0.57)**	22.05(±1.27)
	70%	Gaussian	16.57(±3.08)	3.21(±3.11)	**17.56(±0.72)**	**2.22(±1.17)**	19.78(±0.90)
		Poisson	16.79(±3.21)	3.74(±3.41)	**19.78(±0.87)**	**0.75(±0.25)**	20.53(±0.81)
		Impulse	17.22(±3.58)	3.65(±3.58)	**18.85(±0.88)**	**2.03(±0.68)**	20.88(±1.09)

dB for SURE-ES and 1.07 dB for WMV-ES. Under Poisson noise, ACR-ES achieves the smallest average gap, and under impulse noise CSR-ES gives the best stopping accuracy.

On medical images, CSR-ES is not applicable because it relies on RGB channel consistency. MR-ES and ACR-ES instead provide reference-based stopping signals for grayscale images. The results show that MR-ES is highly effective under impulse noise, while ACR-ES is particularly strong under Gaussian and Poisson noise. For example, on randomly selected images from CT10 and T2-20 under Poisson noise, ACR-ES reduces the average gap to 0.84 dB, compared with 4.34 dB for SURE-ES, 5.38 dB for WMV-ES, and 2.24 dB for MR-ES.

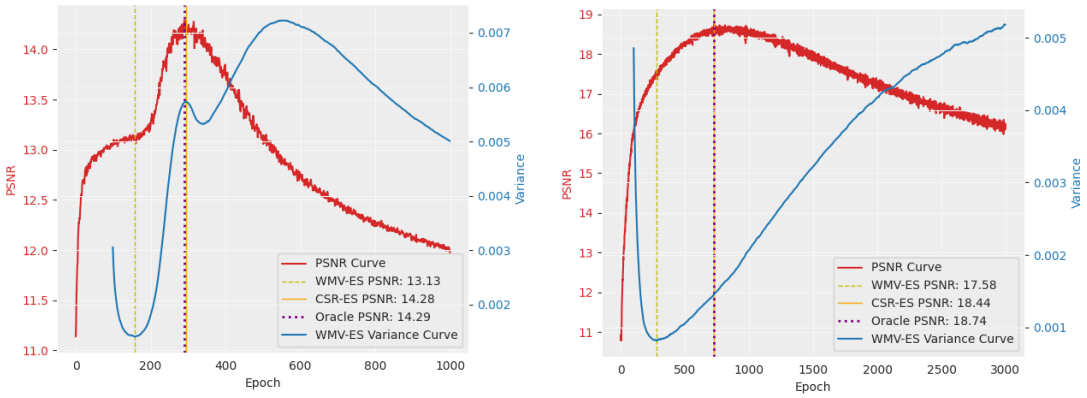
These results show that the proposed criteria do not obtain smaller gaps by sacrificing the attainable oracle quality. Instead, they select stopping points closer to the best attainable point along their corresponding trajectories. The full per-image method-specific oracle comparisons for natural and medical images are provided in Appendix G.3, where the same conclusion holds across individual images and noise types.

4.4 Extension to Super-resolution and Inpainting

Finally, we investigate whether the proposed early-stopping principle extends beyond denoising to other DIP-based inverse problems. We consider super-resolution ($\times 4$ and $\times 8$) and inpainting (50% and 70% missing pixels) on Urban10. We focus on CSR-ES because it can be applied post hoc to a standard DIP trajectory without modifying the optimization objective. We compare CSR-ES with WMV-ES on the same reconstruction trajectories, so both methods share the same PSNR oracle.

Unlike noiseless super-resolution and inpainting, our experiments involve measurements corrupted by Gaussian, Poisson, or impulse noise. Continued optimization may therefore cause DIP to fit measurement noise and artifacts after recovering useful image structure, making early stopping important beyond denoising.

Table 5 reports the results across different degradation operators and noise types. CSR-ES consistently achieves smaller oracle PSNR gaps than WMV-ES. For super-resolution, CSR-ES selects stopping



(a) Super-resolution $\times 8$ with Impulse noise

(b) Inpainting 70% with Gaussian noise

Figure 3: Image restoration results of the WMV and our CSR method.

points very close to the oracle across all tested settings. For inpainting, the gaps are generally larger, but CSR-ES still substantially improves upon WMV-ES.

Figure 3 further illustrates the stopping behavior under super-resolution and inpainting. In the super-resolution example, the WMV-ES variance curve has a local maximum near the PSNR peak, whereas its variance-minimum criterion does not identify the optimal stopping region. In the inpainting example, WMV-ES stops prematurely, well before the PSNR peak. In contrast, CSR-ES selects stopping points closer to the oracle in both cases.

These experiments provide empirical evidence that CSR-ES can be applied to the tested inverse problems beyond denoising. The operator-form analysis in Appendix H establishes a reference-validation relationship in the measurement domain. However, transferring measurement-domain risk control to image-domain reconstruction error requires additional assumptions on the forward operator and the reconstruction class. Thus, the empirical improvements reported here do not constitute a general image-domain performance guarantee.

Ablations and robustness. The appendix provides additional evidence supporting the design choices of the proposed criteria. Appendix G.4 compares MSE-based and bias-based reference scores. The results justify using bias-based ACR-ES for natural color images, where cross-reference bias provides useful information beyond direct MSE consistency, and MSE-based MR-ES for grayscale medical images, where direct mask-reference consistency is more stable. Appendix G.5 studies the auxiliary corrupted references used by ACR-ES. The results show that ACR-ES remains stable under moderate changes of the auxiliary noise magnitude and under mismatched auxiliary noise levels, denoted as $\tau_1 \neq \tau_2$, indicating that it does not require precise tuning of the auxiliary corruption strength. We further evaluate CSR-ES with ASeqDIP and Deep Decoder in Appendix G.6 to test whether stopping remains useful when overfitting is mitigated through an explicit regularizer or architectural underparameterization.

Overall, the experiments show that the proposed reference-based early-stopping criteria select stopping points much closer to the oracle than variance-based stopping across natural images, medical images, and additional inverse problems. CSR-ES is simple and transferable because it is post-hoc and requires no change to DIP training. MR-ES extends the reference-based principle to single-channel medical images where channel consistency is unavailable. ACR-ES provides a more flexible auxiliary-reference signal, especially useful for challenging or signal-dependent corruptions.

5 Conclusion

DIP early stopping can be reduced to a validation problem. To address the absence of clean validation data, we proposed self-referenced criteria that construct pseudo-references from the noisy observation itself. Three concrete methods instantiate this principle: CSR-ES, MR-ES , and ACR-ES. Across experiments under multiple image datasets and several noise types for imaging inverse problems, these criteria consistently select stopping points closer to the oracle than existing baselines.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement No. HR0011-25-2-0021. I.A. was supported by the U.S. Department of Energy through the Laboratory Directed Research and Development at Los Alamos National Laboratory (LANL) under project 20250048DR. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

References

[1] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *ICLR*, pages 1–28, 2024.

[2] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models in medical imaging. *Acta Numerica*, 28:1–174, 2019.

[3] Aviad Levis, Pratul P Srinivasan, Andrew A Chael, Ren Ng, and Katherine L Bouman. Gravitationally lensed black hole emission tomography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[4] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

[5] Hengkang Wang, Xu Zhang, Taihui Li, Yuxiang Wan, Tiancong Chen, and Ju Sun. Dmplug: A plug-in method for solving inverse problems with diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[6] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[7] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Springer, 1996.

[8] Michael T McCann, Kyong Hwan Jin, and Michael Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

[9] Mario Bertero and Patrizia Boccacci. *Introduction to Inverse Problems in Imaging*. Institute of Physics Publishing, 1998.

[10] George Yiasemis, Jan-Jakob Sonke, Clarisa Sánchez, and Jonas Teuwen. Recurrent variational network: A deep learning inverse problem solver applied to the task of accelerated mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 732–741, June 2022.

[11] Jiaming Liu, Yu Sun, Cihat Eldeniz, Weijie Gan, Hongyu An, and Ulugbek S Kamilov. Rare: Image reconstruction using deep priors learned without groundtruth. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1088–1099, 2020.

[12] Hyungjin Chung, Jeongsol Kim, and Jong Chul Ye. Diffusion models for inverse problems. *arXiv preprint arXiv:2508.01975*, 2025.

[13] Amer Essakine, Yanqi Cheng, Chun-Wun Cheng, Lipei Zhang, Zhongying Deng, Lei Zhu, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. Where do we stand with implicit neural representations? a technical and performance survey. *Transactions on Machine Learning Research*.

[14] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[15] Liyue Shen, John Pauly, and Lei Xing. Nerp: Implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[16] Kushal Vyas, Ahmed Imtiaz Humayun, Aniket Dashpute, Richard G Baraniuk, Ashok Veeraraghavan, and Guha Balakrishnan. Learning transferable features for implicit neural representations. 2025.

[17] Luca De Luigi, Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Deep learning on implicit neural representations of shapes. *arXiv preprint arXiv:2302.05438*, 2023.

[18] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022.

[19] Amirali Molaei, Amirhossein Aminimehr, Armin Tavakoli, Amirhossein Kazerouni, Bobby Azad, Reza Azad, and Dorit Merhof. Implicit neural representation in medical imaging: A comparative survey. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2381–2391, 2023.

[20] R Heckel et al. Deep decoder: Concise image representations from untrained non-convolutional networks. In *International Conference on Learning Representations*, page 2736–2750, 2019.

[21] Davide Sapienza, Giorgia Franchini, Elena Govi, Marko Bertogna, and Marco Prato. Deep image prior for medical image denoising, a study about parameter initialization. *Frontiers in Applied Mathematics and Statistics*, 8:995225, 2022.

[22] Hengyue Liang, Taihui Li, and Ju Sun. A baseline method for removing invisible image watermarks using deep image prior. *Transactions on Machine Learning Research (TMLR)*, 2025.

[23] Taihui Li, Hengkang Wang, Zhong Zhuang, and Ju Sun. Deep random projector: Accelerated deep image prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18176–18185, 2023.

[24] Ismail R Alkhouri, Shijun Liang, Evan Bell, Qing Qu, Rongrong Wang, and Saiprasad Ravishankar. Image reconstruction via autoencoding sequential deep image prior. *Advances in Neural Information Processing Systems*, 37:18988–19012, 2024.

[25] Hengkang Wang, Taihui Li, Zhong Zhuang, Tiancong Chen, Hengyue Liang, and Ju Sun. Early stopping for deep image prior. *Transactions on Machine Learning Research*, pages 1–40, 2023.

[26] Ismail Alkhouri, Evan Bell, Avrajit Ghosh, Shijun Liang, Rongrong Wang, and Saiprasad Ravishankar. Understanding untrained deep models for inverse problems: Algorithms and theory. *IEEE Signal Processing Magazine*, 2025.

[27] Shijun Liang, Evan Bell, Qing Qu, Rongrong Wang, and Saiprasad Ravishankar. Analysis of deep image prior and exploiting self-guidance for image reconstruction. *IEEE Transactions on Computational Imaging*, 2025.

[28] Jiaming Liu, Yu Sun, Xiaojian Xu, and Ulugbek S Kamilov. Image restoration using total variation regularized deep image prior. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7715–7719. Ieee, 2019.

[29] Dominique Fourdrinier, William E Strawderman, and Martin T Wells. *Shrinkage estimation*. Springer, 2018.

[30] Alessandro Benfenati, Ambra Catozzi, Giorgia Franchini, and Federica Porta. Early stopping strategies in deep image prior. *Soft Computing*, 29(8):4153–4174, 2025.

[31] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. *International Conference on Learning Representations*, 2019.

Appendix

A Reference Validation Theory

This appendix formalizes the validation principle used in Section 2. Throughout this section, the DIP trajectory $\{\hat{x}_t(y_1)\}_{t=1}^T$ is trained only on

$$y_1 = x + \eta_1,$$

and the reference image is

$$y_2 = x + \eta_2,$$

where η_2 is independent of y_1 and of the fitted trajectory conditional on y_1.

A.1 Setup and notation

For each iteration t, define the clean risk

$$R(t) := \frac{1}{n}\|\hat{x}_t(y_1) - x\|^2$$

and the reference validation loss

$$V_t := \frac{1}{n}\|\hat{x}_t(y_1) - y_2\|^2.$$

Let

$$\hat{t} \in \arg\min_{1\le t\le T} V_t, \qquad t^\star \in \arg\min_{1\le t\le T} R(t).$$

The goal is to compare $R(\hat{t})$ to the oracle risk $R(t^\star)$.

A.2 Single-reference concentration

Theorem A.1 (Single-reference validation)**.** *Assume that the entries of η_2 are independent, zero-mean, sub-Gaussian random variables with variance proxy σ^2 and $\mathbb{E}\eta_{2,i}^2 = \sigma^2$. Then there exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$, uniformly over $1 \le t \le T$,*

$$\left|V_t - R(t) - \sigma^2\right| \le C\left[M_T\sqrt{\frac{\log(T/\delta)}{n}} + \sigma^2\frac{\log(T/\delta)}{n}\right],$$

where

$$M_T := \sigma \sup_{1\le t\le T}\sqrt{R(t)} + \sigma^2.$$

Proof. Condition on y_1 and hence on the entire DIP trajectory $\{\hat{x}_t(y_1)\}_{t=1}^T$. For each fixed t, write

$$a_t := \hat{x}_t(y_1) - x.$$

Since $y_2 = x + \eta_2$,

$$V_t = \frac{1}{n}\|a_t - \eta_2\|^2 = \frac{1}{n}\|a_t\|^2 - \frac{2}{n}\langle a_t, \eta_2\rangle + \frac{1}{n}\|\eta_2\|^2.$$

Therefore

$$V_t - R(t) - \sigma^2 = -\frac{2}{n}\langle a_t, \eta_2\rangle + \frac{1}{n}\left(\|\eta_2\|^2 - n\sigma^2\right).$$

The first term is a centered sub-Gaussian linear functional. Conditional on a_t, standard sub-Gaussian concentration gives, with probability at least $1 - \delta/(2T)$,

$$\left|\frac{2}{n}\langle a_t, \eta_2\rangle\right| \le C\sigma\sqrt{R(t)}\sqrt{\frac{\log(T/\delta)}{n}}.$$

The second term is a centered average of sub-exponential random variables $\eta_{2,i}^2 - \sigma^2$. Bernstein or Hanson–Wright concentration gives, with probability at least $1 - \delta/(2T)$,

$$\left| \frac{1}{n} \left(\|\eta_2\|^2 - n\sigma^2 \right) \right| \leq C\sigma^2 \left[\sqrt{\frac{\log(T/\delta)}{n}} + \frac{\log(T/\delta)}{n} \right].$$

Combining the two inequalities and taking a union bound over $t = 1, \ldots, T$ yields the stated result. □

A.3 Oracle inequality

Corollary A.1.1 (Near-oracle stopping)**.** *Under the event in Theorem A.1, define*

$$\varepsilon_n := C \left[M_T \sqrt{\frac{\log(T/\delta)}{n}} + \sigma^2 \frac{\log(T/\delta)}{n} \right].$$

Then the stopping time $\hat{t} \in \arg\min_t V_t$ *satisfies*

$$R(\hat{t}) \leq R(t^\star) + 2\varepsilon_n.$$

Proof. On the event of Theorem A.1,

$$R(t) \leq V_t - \sigma^2 + \varepsilon_n \quad \text{and} \quad V_t - \sigma^2 \leq R(t) + \varepsilon_n$$

for all t. Hence

$$R(\hat{t}) \leq V_{\hat{t}} - \sigma^2 + \varepsilon_n \leq V_{t^\star} - \sigma^2 + \varepsilon_n \leq R(t^\star) + 2\varepsilon_n.$$

□

A.4 PSNR interpretation

If image intensities are normalized so that the peak value is $I_{\max}$, then

$$\mathrm{PSNR}(t) = 10 \log_{10} \left(\frac{I_{\max}^2}{R(t)} \right).$$

If $R(\hat{t}) \leq R(t^\star) + 2\varepsilon_n$, then

$$\mathrm{PSNR}(t^\star) - \mathrm{PSNR}(\hat{t}) = 10 \log_{10} \left(\frac{R(\hat{t})}{R(t^\star)} \right) \leq 10 \log_{10} \left(1 + \frac{2\varepsilon_n}{R(t^\star)} \right).$$

Thus the PSNR loss is small whenever the oracle MSE is bounded away from zero and the validation error ε_n is small.

B Pseudo-Reference Validation

This appendix justifies the pseudo-reference discussion in Section 2. The purpose is not to claim that a pseudo-reference is fully independent, but to identify the bias introduced by the component shared between fitting and validation.

B.1 Risk estimated by pseudo-validation

Assume that the observed image and pseudo-reference can be decomposed as

$$y = x + s + \eta, \qquad \tilde{y} = x + s + \tilde{\eta},$$

where s is shared between fitting and validation, while $\tilde{\eta}$ is zero-mean and independent of the DIP trajectory conditional on y and s. Define

$$\tilde{V}_t := \frac{1}{n} \|\hat{x}_t(y) - \tilde{y}\|^2.$$

Proposition B.1 (Pseudo-validation estimates effective-target risk)**.** *Conditioned on y and s,*

$$\mathbb{E}_{\tilde{\eta}}\Big[\tilde{V}_t \mid y, s\Big] = \frac{1}{n}\|\hat{x}_t(y) - (x+s)\|^2 + \frac{1}{n}\mathbb{E}\|\tilde{\eta}\|^2.$$

In particular, if $\tilde{\eta}$ has per-pixel variance $\sigma_{\tilde{\eta}}^2$, then

$$\mathbb{E}_{\tilde{\eta}}\Big[\tilde{V}_t \mid y, s\Big] = \frac{1}{n}\|\hat{x}_t(y) - (x+s)\|^2 + \sigma_{\tilde{\eta}}^2.$$

Proof. Let

$$b_t := \hat{x}_t(y) - (x+s),$$

then

$$\tilde{V}_t = \frac{1}{n}\|b_t - \tilde{\eta}\|^2 = \frac{1}{n}\|b_t\|^2 - \frac{2}{n}\langle b_t, \tilde{\eta}\rangle + \frac{1}{n}\|\tilde{\eta}\|^2.$$

Conditioned on y and s, b_t is fixed and $\mathbb{E}\tilde{\eta} = 0$, so the cross term has mean zero. This proves the claim. □

Proposition B.1 gives a conservative interpretation: pseudo-validation tracks risk to the effective target $x + s$, not necessarily to the clean image x. Therefore, the construction of $\tilde{y}$ should make the shared component s small or ensure that s is not preferentially fitted before the remaining noise.

B.2 Average-case behavior under DIP noise fitting

The worst-case in Proposition B.1 is often overly pessimistic in DIP for the following reason. In practice, the shared component s and the non-shared component η may have the same statistical structure, e.g., both are Gaussian-like noise components. In this case, DIP has no mechanism to preferentially fit s before fitting η. Instead, the typical optimization trajectory first captures the structured image component x, and only later gradually fits the entire noise component $s + \eta$.

Assume that, after the oracle stopping region, the DIP trajectory gradually fits the noise in $y = x+s+\eta$ according to

$$\hat{x}_t(y) \approx x + \alpha_t(s+\eta), \qquad \alpha_t \in [0,1].$$

Assume also that η and $\tilde{\eta}$ are independent, zero-mean, and have the same per-pixel variance σ_η^2. Let

$$S := \frac{1}{n}\|s\|^2.$$

Then

$$\hat{x}_t(y) - \tilde{y} \approx (\alpha_t - 1)s + \alpha_t\eta - \tilde{\eta}.$$

Taking expectation over the non-shared noises conditional on s gives

$$\mathbb{E}[\tilde{V}_t \mid s] \approx (1-\alpha_t)^2 S + \alpha_t^2\sigma_\eta^2 + \sigma_\eta^2 = (1-\alpha_t)^2 S + (\alpha_t^2 + 1)\sigma_\eta^2.$$

Proposition B.2 (Effect of shared noise in the stylized model)**.** *As a function of $\alpha \in [0,1]$, the expression*

$$(1-\alpha)^2 S + (\alpha^2+1)\sigma_\eta^2$$

is minimized at

$$\alpha^\star = \frac{S}{S + \sigma_\eta^2}.$$

Proof. Differentiate:

$$\frac{d}{d\alpha}\left[(1-\alpha)^2 S + (\alpha^2+1)\sigma_\eta^2\right] = 2(\alpha-1)S + 2\alpha\sigma_\eta^2.$$

Setting the derivative to zero gives

$$\alpha(S + \sigma_\eta^2) = S,$$

and therefore

$$\alpha^\star = \frac{S}{S + \sigma_\eta^2}.$$

The second derivative is $2S + 2\sigma_\eta^2 > 0$, so this is the minimizer. □

Quadratic improvement over the worst case. Suppose the pseudo-reference is constructed so that the shared energy is a $1/m$ fraction of the non-shared variance, $S = \sigma_\eta^2/m$ with $m \geq 1$. Proposition B.2 then gives $\alpha^\star = 1/(m+1)$, and the expected MSE to the clean image x at the model-implied stopping time is

$$\mathbb{E}\Big[\tfrac{1}{n}\|\hat{x}_{t^\star}(y) - x\|^2 \,\Big|\, s\Big] \;\approx\; (\alpha^\star)^2\,(S + \sigma_\eta^2) \;=\; \frac{\sigma_\eta^2}{m(m+1)} \;=\; \mathcal{O}(1/m^2),$$

which is quadratically tighter than the worst-case excess $\|s\|^2/n = \sigma_\eta^2/m = \mathcal{O}(1/m)$ implied by Proposition B.1. Modest decorrelation (m moderately large) is therefore sufficient for near-oracle MSE under the noise-fitting model.

B.3 Interpretation and limitations

The formula for $\alpha^\star$ explains the role of pseudo-reference design. If $S \ll \sigma_\eta^2$, then $\alpha^\star \approx 0$, so the validation curve is minimized near the beginning of the noise-fitting phase. This corresponds to a useful early-stopping signal. If $S \gg \sigma_\eta^2$, then $\alpha^\star \approx 1$, so the validation minimum may occur late, after DIP has already fitted much of the noise.

This model is only a stylized description of DIP dynamics. It assumes that the shared and non-shared components have comparable statistical structure (e.g. both are Gaussian noise), and that the network fits them at similar rates. It should therefore be read as a design principle rather than a complete dynamical theory of DIP optimization.

C Algorithmic Details

This appendix records the pseudocode for the three pseudo-reference early stopping criteria introduced in Section 3. The main text gives the defining validation curves; the algorithms below make explicit the inputs, selected stopping time, and returned reconstruction.

Algorithm 1 Channel-Similarity Reference Early Stopping (CSR-ES)

Require: Noisy RGB image $y = (y_R, y_G, y_B)$; DIP trajectory $\{\hat{x}_t\}_{t=1}^T$

1: Compute pairwise channel distances

$$d_{ij} = \frac{1}{n}\|y_i - y_j\|^2, \qquad i, j \in \{R, G, B\},\ i \neq j.$$

2: Select the closest channel pair

$$(i^\star, j^\star) = \arg\min_{i \neq j} d_{ij}.$$

3: Evaluate the cross-channel validation curve

$$V_t^{\mathrm{CSR}} = \frac{1}{n}\|\hat{x}_{t,i^\star} - y_{j^\star}\|^2.$$

4: Select

$$\hat{t}_{\mathrm{CSR}} = \arg\min_{1 \leq t \leq T} V_t^{\mathrm{CSR}}.$$

5: **return** $\hat{x}_{\hat{t}_{\mathrm{CSR}}}$.

Unlike MSE-based stopping rules that search for a minimum, the spectral-bias ACR-ES Algorithm 3 uses an $\arg\max$ criterion. This is because, due to the spectral bias of DIP, the average-frequency score often follows an increase-then-decrease pattern during DIP training. In early iterations, the network mainly reconstructs low-frequency structures; as training progresses, meaningful higher-frequency details are recovered and the score increases. After the near-oracle region, further optimization tends to fit noise or unstable artifacts, reducing the reference-consistent spectral concentration. Thus, the peak of the spectral-bias score marks the transition from useful detail recovery to overfitting and provides an effective stopping indicator.

Algorithm 2 Mask-Reference Early Stopping (MR-ES)

Require: Noisy image y; retention probability p; DIP iterations T

1: Sample or choose a mask $M \in \{0,1\}^n$ with high retention rate, e.g., $\mathbb{P}(M_i = 1) = p = 0.98$, and set $H = 1 - M$.

2: Train DIP using the masked loss

$$\min_{\theta} \|M \odot (f_\theta(z) - y)\|^2.$$

3: Along the resulting trajectory $\{\hat{x}_t\}_{t=1}^T$, evaluate only on held-out pixels:

$$V_t^{\mathrm{MR}} = \frac{1}{\sum_i H_i} \|H \odot (\hat{x}_t - y)\|^2.$$

4: Select

$$\hat{t}_{\mathrm{MR}} = \arg\min_{1 \le t \le T} V_t^{\mathrm{MR}}.$$

5: **return** $\hat{x}_{\hat{t}_{\mathrm{MR}}}$.

Algorithm 3 Augmented-Channel Reference Early Stopping (ACR-ES)

Require: Noisy image y with C channels; auxiliary noise scale τ; burn-in t_0; DIP iterations T

1: Generate two independent corrupted measurements

$$y^{(1)} = \mathrm{Corrupt}(y; \zeta_1), \qquad y^{(2)} = \mathrm{Corrupt}(y; \zeta_2),$$

where ζ_1 and ζ_2 with noise level τ are user-specified auxiliary independent perturbations following the same corruption model.

2: Construct the augmented fitting target

$$Y = \mathrm{concat}(y, y^{(1)}).$$

3: Train a DIP network with $2C$ output channels to fit Y, producing outputs $\hat{X}_t = f_{\theta_t}(z)$.

4: Let $\hat{X}_{t,\mathrm{aux}}$ denote the auxiliary output channels $C+1 : 2C$. Evaluate

$$B_t^{\mathrm{ACR}} = \frac{\sum_\omega \|\omega\|^2 \left| \mathcal{F}(\hat{X}_{t,\mathrm{aux}} - y^{(2)})(\omega) \right|^2}{\sum_\omega \left| \mathcal{F}(\hat{X}_{t,\mathrm{aux}} - y^{(2)})(\omega) \right|^2}.$$

5: Select

$$\hat{t}_{\mathrm{ACR}} = \arg\max_{t \ge t_0} B_t^{\mathrm{ACR}}.$$

6: **return** the primary reconstruction $\hat{x}_{\hat{t}_{\mathrm{ACR}}} = \hat{X}_{\hat{t}_{\mathrm{ACR}}, 1:C}$.

D Theory for Channel-Similarity Reference

This appendix records the algorithm-specific justification for CSR-ES used in Section 3. We assume N_R, N_G, N_B are zero-mean, mutually independent, and sub-Gaussian with variance proxy σ^2. Let $\hat{x}_{t,c}$ denote the c-channel of the DIP reconstruction at iteration t. The cross-channel validation curve is

$$V_t^{\mathrm{RGB}} := \tfrac{1}{n} \|\hat{x}_{t,i^\star} - y_{j^\star}\|^2.$$

Writing $y_c = x_c + N_c$, define

$$\Delta := x_{j^\star} - x_{i^\star}, \qquad R_t := \frac{1}{n} \|\hat{x}_{t,i^\star} - x_{i^\star}\|^2, \qquad \rho := \sqrt{\|\Delta\|^2/n}.$$

An unconditional controlled-bias guarantee. Substituting $y_{j^\star} = x_{i^\star} + \Delta + N_{j^\star}$, $\hat{x}_{t,i^\star} - y_{j^\star} = r_t - \Delta - N_{j^\star}$ where $r_t := \hat{x}_{t,i^\star} - x_{i^\star}$. Squaring and dividing by n,

$$V_t^{\mathrm{RGB}} \;=\; R_t + \rho^2 + \tfrac{1}{n}\|N_{j^\star}\|^2 \;-\; \tfrac{2}{n}\langle r_t, \Delta\rangle \;-\; \tfrac{2}{n}\langle r_t, N_{j^\star}\rangle \;+\; \tfrac{2}{n}\langle \Delta, N_{j^\star}\rangle.$$

The first term is the reconstruction risk we wish to monitor; the next two are t-independent constants (in expectation, for the latter); the remaining three cross terms determine whether V_t^{RGB} tracks R_t up to a constant offset.

Proposition D.1 (Controlled-bias guarantee, uniform in t)**.** *With probability at least $1-\delta$ over $N_{j^\star}$, simultaneously for all $1 \le t \le T$,*

$$\begin{aligned}\left| V_t^{\mathrm{RGB}} - \left(R_t + \rho^2 + \sigma^2\right) \right| \;\le\; & 2\rho\sqrt{R_t} \;+\; \left|\tfrac{2}{n}\langle r_t, N_{j^\star}\rangle\right| \\ & +\; C\left[\rho\sigma\sqrt{\frac{\log(T/\delta)}{n}} \;+\; \sigma^2\left(\sqrt{\frac{\log(T/\delta)}{n}} + \frac{\log(T/\delta)}{n}\right)\right].\end{aligned}$$

The first term comes from Cauchy–Schwarz on $\langle r_t, \Delta\rangle$; the second is the residual cross-correlation between the reconstruction error and the validation noise; the bracketed terms come from sub-Gaussian concentration of $\langle \Delta, N_{j^\star}\rangle$ and Hanson–Wright concentration of $\|N_{j^\star}\|^2 - n\sigma^2$, with a union bound over $t \in \{1, \dots, T\}$.

When ρ is small and the residual cross-correlation $\frac{1}{n}\langle r_t, N_{j^\star}\rangle$ remains small uniformly in t, the validation curve V_t^{RGB} tracks R_t up to a t-independent additive offset $\rho^2 + \sigma^2$, with concentration error vanishing in n. This justifies the channel-similarity selection step in Algorithm 1: minimizing ρ shrinks both the constant offset and the trajectory-dependent distortion $2\rho\sqrt{R_t}$.

Data-dependent channel-pair selection. The pair $(i^\star, j^\star)$ in Algorithm 1 is selected by minimizing $\frac{1}{n}\|y_i - y_j\|^2$ over pairs, which expands as

$$\tfrac{1}{n}\|y_i - y_j\|^2 \;=\; \tfrac{1}{n}\|x_i - x_j\|^2 \;+\; \tfrac{1}{n}\|N_i - N_j\|^2 \;+\; \tfrac{2}{n}\langle x_i - x_j, N_i - N_j\rangle.$$

The selection therefore depends on the noise realizations through the second and third terms, with typical bias of order $\sigma^2/\sqrt{n}$ arising from conditioning on the rank statistic among three pairs. In what follows we condition on the selected pair $(i^\star, j^\star)$ and treat Δ as deterministic; this is equivalent to ignoring the $O(\sigma^2/\sqrt{n})$ adaptive-selection bias, which is dominated by ρ^2 and σ^2 in the regimes of interest. A fully rigorous treatment would replace expectations over N_c with conditional expectations given the selection event, at the cost of additional notation.

An idealized argmin recovery model. Under additional assumptions, the residual cross-correlation vanishes in expectation and the rule recovers the per-channel argmin during the noise-fitting phase of training.

Assumption D.2 (Reference-channel holdout)**.** *The reconstruction $\hat{x}_{t,i^\star}$ is independent of $N_{j^\star}$.*

Assumption D.3 (Per-channel overfitting trajectory)**.** *There exists t_0 such that for $t \ge t_0$,*

$$\hat{x}_{t,i^\star} \;\approx\; x_{i^\star} + \alpha_t\, N_{i^\star}, \qquad \alpha_t \in [0,1],$$

with α_t monotonically increasing in t.

Assumption D.2 does *not* hold for standard full-RGB DIP, since the network parameters θ_t are updated using gradients from all three channels and therefore $\hat{x}_{t,i^\star}$ depends on $N_{j^\star}$ through shared weights. The assumption holds exactly if the $j^\star$-channel is excluded from the training loss (see Remark D.5 below), and approximately when cross-channel gradient leakage is small relative to the risk variation that separates the desired stopping region from later overfitted iterates. Empirically, the effectiveness of CSR-ES suggests this term is not dominant in the tested regimes; we do not attempt to prove such a bound here. Assumption D.3 is the per-channel analogue of the trajectory model used in the shared-noise analysis.

Theorem D.4 (Idealized risk tracking in the noise-fitting phase)**.** *Under Assumptions D.2 and D.3, for $t \ge t_0$,*

$$\mathbb{E}\left[V_t^{\mathrm{RGB}}\right] \;=\; \alpha_t^2\sigma^2 + \sigma^2 + \rho^2, \qquad \mathbb{E}[R_t] \;=\; \alpha_t^2\sigma^2.$$

Both quantities are strictly monotone increasing in α_t, and so on $\{t \ge t_0\}$ they are minimized at the same iterate.

Proof. Under Assumption D.2, $\mathbb{E}\langle r_t, N_{j^\star}\rangle = 0$ since $N_{j^\star}$ is mean zero and independent of the first factor. Under Assumption D.3, $\mathbb{E}\langle r_t, \Delta\rangle = \alpha_t \mathbb{E}\langle N_{i^\star}, \Delta\rangle = 0$ since Δ depends only on clean channels and $N_{i^\star}$ is mean zero. Furthermore $\mathbb{E}\langle \Delta, N_{j^\star}\rangle = 0$ and $\mathbb{E}\|N_{j^\star}\|^2 = n\sigma^2$. Taking expectation in (1) gives $\mathbb{E}[V_t^{\mathrm{RGB}}] = \mathbb{E}[R_t] + \rho^2 + \sigma^2$, with $\mathbb{E}[R_t] = \alpha_t^2\sigma^2$ from Assumption D.3. □

Scope of Theorem D.4. Two limitations deserve emphasis. First, both $\mathbb{E}[V_t^{\mathrm{RGB}}]$ and $\mathbb{E}[R_t]$ are monotone increasing on $\{t \geq t_0\}$, so the theorem describes the noise-fitting phase only. The early-stopping mechanism arises from combining this with the descending phase of R_t during image fitting, where the trajectory model does not apply: V_t^{RGB} is non-monotone across the full trajectory, decreasing during image fitting and increasing during noise fitting, with the rule selecting the transition. Theorem D.4 establishes that the rule does not drift past the transition, but does not by itself identify the transition.

Second, the theorem requires the idealized holdout (Assumption D.2) which is not satisfied by standard full-RGB DIP. The implemented method therefore corresponds to an approximate version of the theorem in which the residual cross-correlation $\frac{1}{n}\langle r_t, N_{j^\star}\rangle$ is treated as small but nonzero. The unconditional Proposition D.1 applies in either case and provides the nonparametric guarantee that channel-similarity selection serves to tighten.

Remark D.5 (A holdout variant of the algorithm)**.** *The idealized model becomes exact for an explicit holdout variant in which the $j^\star$-channel is removed from the training loss:*

$$\theta_t \;\leftarrow\; \theta_{t-1} - \eta\,\nabla_\theta \sum_{c \neq j^\star} \tfrac{1}{n}\|\hat{x}_{t-1,c} - y_c\|^2.$$

This variant has the cleanest theoretical justification but trains on only two of the three channels, which may degrade reconstruction quality. The implemented full-RGB algorithm is the practical compromise; empirical comparison is left to future work.

Remark D.6 (Departures from per-channel noise independence)**.** *Real RGB noise is approximately but not exactly per-channel independent; demosaicing and sensor characteristics can induce mild cross-channel correlation. If $\mathbb{E}\langle N_{i^\star}, N_{j^\star}\rangle = \rho_{ij} n\sigma^2$, the cross term in Proposition D.1 acquires an additional contribution of order $\alpha_t \rho_{ij}\sigma^2$, which shifts the argmin only when $|\rho_{ij}|$ is sizeable. For typical natural images and standard sensor models, $|\rho_{ij}|$ is small enough that the present analysis remains a useful approximation.*

E Mask-Reference Validation

This appendix gives a partial justification for the mask-reference early stopping criterion used by MR-ES. The key point is that, if held-out pixels are excluded from the training loss, then the noise at those held-out pixels is independent of the fitted trajectory at those locations.

E.1 Setup

Let

$$y = x + \xi,$$

where $\xi \in \mathbb{R}^n$ has independent, zero-mean, sub-Gaussian entries with variance proxy σ^2 and per-pixel variance σ^2. Let $M \in \{0,1\}^n$ be a mask independent of ξ, and let

$$H := 1 - M$$

be the held-out indicator. Define

$$m := \sum_{i=1}^{n} H_i.$$

Assume that the DIP trajectory is trained using only the retained pixels:

$$\theta_t \approx \arg\min_\theta \|M \odot (f_\theta(z) - y)\|^2 .$$

Thus the held-out pixels do not participate in fitting. Define the held-out validation loss

$$V_t^{\mathrm{mask}} := \frac{1}{m}\|H \odot (\hat{x}_t - y)\|^2$$

and the held-out clean risk

$$R_H(t) := \frac{1}{m}\|H \odot (\hat{x}_t - x)\|^2 .$$

E.2 Held-out independence

Lemma E.1 (Independence at held-out positions)**.** *For every pixel i with $H_i = 1$, the noise entry ξ_i is independent of $\hat{x}_t(i)$.*

Proof. The training loss uses only pixels j with $M_j = 1$. Hence, conditional on the fixed input z and the mask M, the parameters θ_t are functions only of the retained observations $\{y_j : M_j = 1\}$, equivalently of $\{x_j + \xi_j : M_j = 1\}$. If $H_i = 1$, then $M_i = 0$, so ξ_i is not used in the training loss. Since the noise entries are independent and the mask is independent of the noise, ξ_i is independent of the retained noise variables and hence independent of θ_t. Therefore ξ_i is independent of $\hat{x}_t(i) = f_{\theta_t}(z)_i$. □

E.3 Held-out validation theorem

Theorem E.2 (Validation on the held-out region)**.** *Under the setup above, there exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$, uniformly over $1 \le t \le T$,*

$$\left|V_t^{\mathrm{mask}} - R_H(t) - \sigma^2\right| \le C\left[M_H\sqrt{\frac{\log(T/\delta)}{m}} + \sigma^2\frac{\log(T/\delta)}{m}\right],$$

where

$$M_H := \sigma \sup_{1\le t\le T}\sqrt{R_H(t)} + \sigma^2.$$

Consequently, if

$$\hat{t} \in \arg\min_t V_t^{\mathrm{mask}}, \qquad t_H^\star \in \arg\min_t R_H(t),$$

then

$$R_H(\hat{t}) \le R_H(t_H^\star) + 2C\left[M_H\sqrt{\frac{\log(T/\delta)}{m}} + \sigma^2\frac{\log(T/\delta)}{m}\right].$$

Proof. Using $y = x + \xi$,

$$V_t^{\mathrm{mask}} = \frac{1}{m}\|H \odot (\hat{x}_t - x - \xi)\|^2.$$

Expanding gives

$$V_t^{\mathrm{mask}} - R_H(t) - \sigma^2 = -\frac{2}{m}\langle H \odot (\hat{x}_t - x), \xi\rangle + \frac{1}{m}\left(\|H \odot \xi\|^2 - m\sigma^2\right).$$

By Lemma E.1, the held-out noise entries are independent of the fitted trajectory at the held-out pixels. Therefore the cross term is a sub-Gaussian linear functional conditional on the trajectory and mask, with deviation of order

$$\sigma\sqrt{R_H(t)}\sqrt{\frac{\log(T/\delta)}{m}}.$$

The quadratic term is controlled by standard sub-exponential concentration for squares of sub-Gaussian variables, with deviation of order

$$\sigma^2\left[\sqrt{\frac{\log(T/\delta)}{m}} + \frac{\log(T/\delta)}{m}\right].$$

A union bound over $t = 1, \dots, T$ yields the uniform concentration inequality. The oracle inequality follows by the same comparison argument as in Corollary A.1.1. □

E.4 Transfer to full-image risk

Theorem E.2 controls the held-out risk $R_H(t)$, whereas the desired reconstruction metric is usually the full-image risk

$$R(t) := \frac{1}{n}\|\hat{x}_t - x\|^2.$$

A transfer condition is therefore needed.

Corollary E.2.1 (Transfer from held-out risk to full-image risk)**.** *Suppose that the held-out region is representative of the full image along the trajectory in the sense that*

$$\sup_{1\le t\le T} |R_H(t) - R(t)| \le \kappa.$$

Then the MR-ES stopping time satisfies

$$R(\hat{t}) \le \min_{1\le t\le T} R(t) + 2\kappa + 2C\left[M_H\sqrt{\frac{\log(T/\delta)}{m}} + \sigma^2\frac{\log(T/\delta)}{m}\right].$$

Proof. Let

$$t^\star \in \arg\min_t R(t),$$

by the representativeness assumption,

$$R(\hat{t}) \le R_H(\hat{t}) + \kappa.$$

By Theorem E.2,

$$R_H(\hat{t}) \le R_H(t_H^\star) + 2\varepsilon_m \le R_H(t^\star) + 2\varepsilon_m,$$

where

$$\varepsilon_m = C\left[M_H\sqrt{\frac{\log(T/\delta)}{m}} + \sigma^2\frac{\log(T/\delta)}{m}\right].$$

Again using representativeness,

$$R_H(t^\star) \le R(t^\star) + \kappa.$$

Combining the three inequalities gives the result. □

E.5 When is the transfer term small?

For random masks with independent Bernoulli entries and with the mask independent of the noise, $R_H(t)$ is an unbiased estimate of $R(t)$ over the randomness of the mask:

$$\mathbb{E}_M[R_H(t)] = R(t),$$

up to the usual normalization by the random held-out size. Concentration over the mask then suggests that $R_H(t)$ is close to $R(t)$ when the held-out set is large and representative. The mask ratio therefore creates a trade-off: retaining more pixels preserves the DIP training signal, but holding out fewer pixels increases the variance of the validation estimate.

E.6 Implementation requirement

The validation interpretation requires that held-out pixels do not enter the training loss. Thus MR-ES should train with

$$\|M \odot (f_\theta(z) - y)\|^2,$$

or an equivalent loss that ignores held-out pixels. It should not train against the zero-filled image through

$$\|f_\theta(z) - M \odot y\|^2,$$

because that objective explicitly supervises the network to output zero at held-out positions. In that case the held-out pixels are no longer unused, and Lemma E.1 no longer supports the validation interpretation.

E.7 Failure modes

The MR-ES justification has two main limitations. First, for signal-dependent noise such as Poisson noise, the validation offset is no longer a constant σ^2 across pixels. Instead, the noise variance depends on the underlying intensity. If the held-out region has a different intensity distribution from the full image, the held-out validation curve may be biased.

Second, deterministic or structured masks, such as masks that mostly select background regions in medical images, require a domain-specific representativeness assumption. The held-out validation theorem may still hold on the selected region, but the transfer from $R_H(t)$ to the full-image risk $R(t)$ depends on the size of κ.

F Augmented-Channel Reference Remarks

ACR-ES follows the pseudo-reference principle of Appendix B, but uses an auxiliary output channel rather than a spatial or color-channel split. Given

$$y^{(1)} = y + \zeta_1, \qquad y^{(2)} = y + \zeta_2,$$

the two pseudo-copies share the original observation y but contain independent added perturbations. If $y = x + \xi$, then

$$y^{(1)} = x + \xi + \zeta_1, \qquad y^{(2)} = x + \xi + \zeta_2.$$

Thus ACR-ES can be viewed as an instance of the pseudo-reference decomposition in Appendix B with

$$s = \xi, \qquad \eta = \zeta_1, \qquad \tilde{\eta} = \zeta_2.$$

The shared component is the original observation noise, while the non-shared components are the added auxiliary perturbations. Increasing the auxiliary perturbation scale increases the non-shared component between fitting and validation, while an excessively large perturbation can interfere with the primary reconstruction task. Thus the auxiliary noise level controls the same shared-versus-independent trade-off described in Appendix B.

The frequency-bias score used by ACR-ES should be viewed as a trajectory diagnostic rather than an unbiased risk estimator. It is motivated by the spectral bias of DIP: as training moves from fitting image structure to fitting noise or artifacts, the spectral distribution of the auxiliary residual changes. The empirical ablations in Section 4 compare this score with direct MSE-based alternatives.

G Additional Experimental Results

This section provides additional experimental evidence supporting the main results in Section 4. The main paper focuses on four representative quantitative studies and three visualizations. Here, we provide complementary results that clarify the behavior of the proposed early-stopping criteria from five aspects: (i) analysis of the RGB channel-reference selection used by CSR-ES, (ii) additional trajectory and qualitative visualizations, (iii) method-specific oracle comparisons on natural and medical images, (iv) ablations of MSE-based and bias-based reference scores, (v) sensitivity of ACR-ES to auxiliary noise levels, and (vi) Early stopping beyond vanilla DIP.

G.1 Analysis of RGB Channel References in CSR-ES

CSR-ES relies on the observation that different RGB channels of a natural image often share correlated structures. The method automatically selects a pseudo-reference channel pair from the noisy observation and uses the corresponding cross-channel consistency curve as a validation signal. In this section, we examine all possible RGB channel-reference pairs to verify whether the automatically selected pair is reliable.

Table 6 reports the CSR-ES results over all RGB channel-reference pairs. Rows correspond to the pseudo-reference channel y_j, while columns correspond to the reconstructed channel $\hat{x}_{t,i}$. Each entry shows the stopping iteration selected by the validation curve

$$\|\hat{x}_{t,i} - y_j\|^2/n$$

Table 6: CSR-ES selection over RGB channel pairs. Rows correspond to the RGB channels of y, and columns correspond to the RGB channels of $\hat{x}$. Each cell reports **PSNR** on the first line and the corresponding iteration (iter.) on the second line. The selected channel pair is highlighted in blue.

Gaussian noise

	$\hat{x}_R$	$\hat{x}_G$	$\hat{x}_B$
y_R	–	**20.70** iter. 2666	**20.84** iter. 3139
y_G	**21.22** **iter. 2373**	–	**20.84** iter. 3139
y_B	**15.12** iter. 32	**20.89** iter. 3482	–

Closest: (R, G)
Selected: $(y_G, \hat{x}_R)$ / 21.22 / iter. 2373
Oracle: 21.22 / iter. 2373

Poisson noise

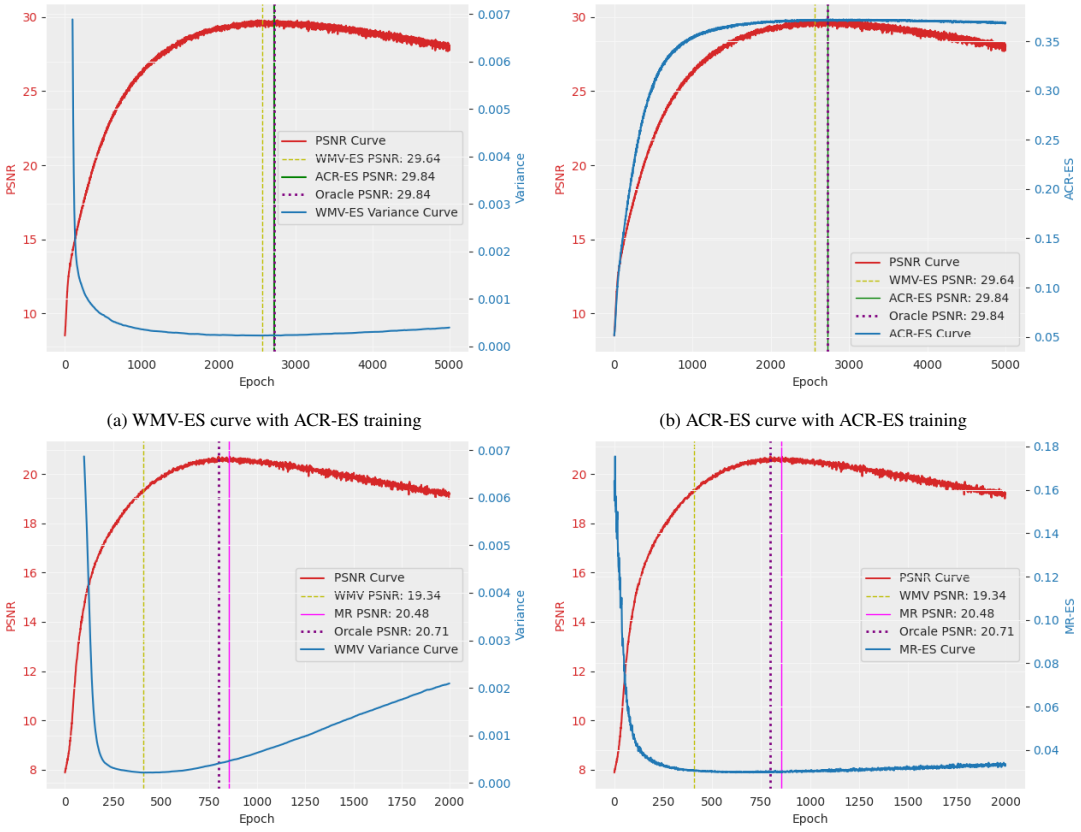
	$\hat{x}_R$	$\hat{x}_G$	$\hat{x}_B$
y_R	–	**22.85** **iter. 4980**	**22.78** iter. 4988
y_G	**22.75** iter. 4925	–	**22.75** iter. 4925
y_B	**22.72** iter. 4937	**22.73** iter. 4974	–

Closest: (R, G)
Selected: $(y_R, \hat{x}_G)$ / 22.85 / iter. 4980
Oracle: 22.85 / iter. 4980

Impulse noise

	$\hat{x}_R$	$\hat{x}_G$	$\hat{x}_B$
y_R	–	**15.48** iter. 20	**15.48** iter. 20
y_G	**23.47** iter. 3036	–	**23.37** iter. 2588
y_B	**23.47** iter. 3036	**23.58** **iter. 2348**	–

Closest: (G, B)
Selected: $(y_B, \hat{x}_G)$ / 23.58 / iter. 2348
Oracle: 23.62 / iter. 2414



(a) WMV-ES curve with ACR-ES training

(b) ACR-ES curve with ACR-ES training

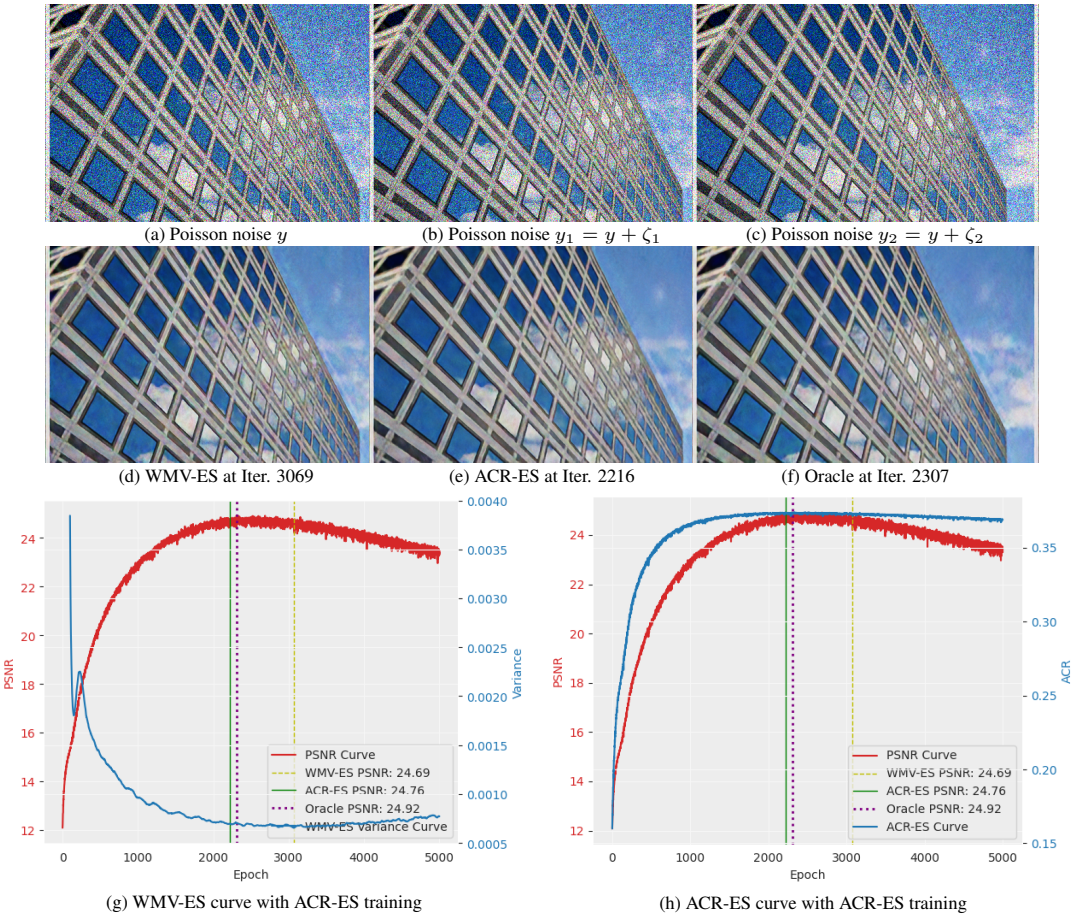
(c) WMV-ES curve with MR-ES training

(d) MR-ES curve with MR-ES training

Figure 4: Early stop for image denoising. First row is the nature image from Set18 corrupted by Poisson noise, the second row is the CT image from CT10 corrupted by Gaussian noise.

and the corresponding reconstruction PSNR. The highlighted entry denotes the pair automatically selected by CSR-ES using the noisy observation only. The oracle result is reported only for evaluation and is not used by CSR-ES.

For Gaussian noise, CSR-ES selects the pair $(\hat{x}_R, y_G)$ and stops at iteration 2373, achieving 21.2249 dB, exactly matching the oracle-best PSNR. For Poisson noise, CSR-ES selects $(\hat{x}_G, y_R)$ and stops at iteration 4980, again exactly matching the oracle-best PSNR of 22.8460 dB. For impulse noise, CSR-ES selects $(\hat{x}_G, y_B)$ and stops at iteration 2348, achieving 23.5807 dB, only 0.0427 dB below the oracle maximum of 23.6234 dB. These results demonstrate that the channel similarity reference provides a reliable pseudo-validation signal for early stopping without access to the clean image.



(a) Poisson noise y (b) Poisson noise $y_1 = y + \zeta_1$ (c) Poisson noise $y_2 = y + \zeta_2$

(d) WMV-ES at Iter. 3069 (e) ACR-ES at Iter. 2216 (f) Oracle at Iter. 2307

(g) WMV-ES curve with ACR-ES training (h) ACR-ES curve with ACR-ES training

Figure 5: Image denoising results on Poisson noise. Comparison of WMV-ES and our ACR-ES.

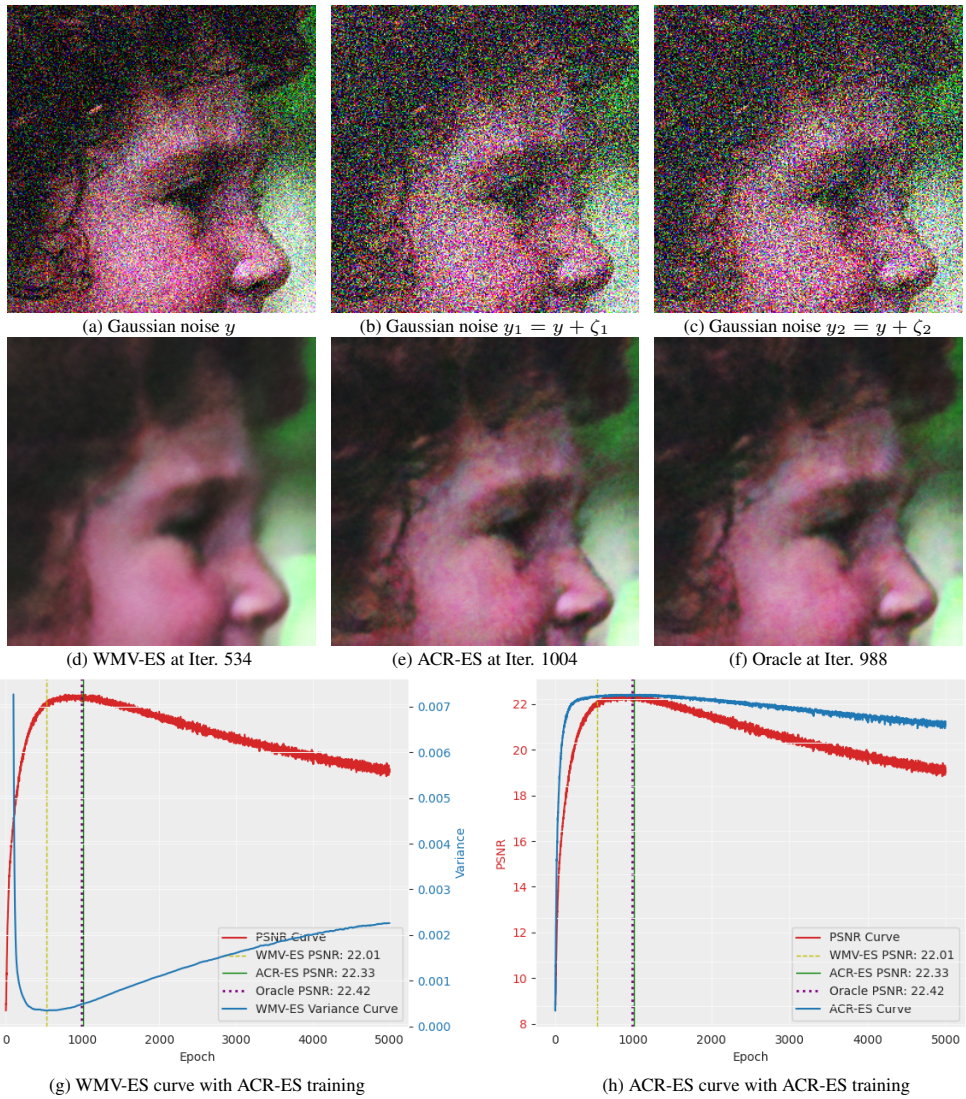
G.2 Additional Trajectory and Qualitative Results

The main paper includes representative trajectory plots showing that the proposed criteria select iterations close to the oracle region. Here, we provide additional curves and qualitative examples to show that this behavior is consistent across different image types and corruption models.

Figure 4 compares WMV-ES with ACR-ES and MR-ES on representative natural and medical images. The WMV curve may reach its minimum before the PSNR peak, leading to premature stopping. In contrast, the proposed reference-based scores better track the region where the reconstruction quality is close to the oracle.

Figures 5 and 6 provide qualitative denoising examples under Poisson and Gaussian noise. Each figure shows the noisy observation, the auxiliary corrupted references, the reconstruction selected by WMV-ES, the reconstruction selected by ACR-ES, and the oracle reconstruction. The selected iteration of ACR-ES is closer to the oracle iteration, and the corresponding reconstruction is visually closer to the oracle-quality result.

Figure 7 shows the denoising results of CT image corrupted by the Gaussian noise. We provide the noisy observation, two auxiliary corrupted references, the reconstruction selected by WMV-ES, the reconstruction selected by ACR-ES, and the oracle reconstruction. The corresponding WMV-ES and ACR-ES curves are also given. The selected iteration of ACR-ES is closer to the oracle iteration. Besides, we also present the denoising results of T2 image corrupted by the Impulse noise in Figure 8. We show the noisy image, the reconstruction selected by WMV-ES, the reconstruction selected by MR-ES, and the oracle reconstruction. The corresponding WMV-ES and MR-ES curves are also given. The selected iteration of MR-ES is closer to the oracle iteration.



(a) Gaussian noise y (b) Gaussian noise $y_1 = y + \zeta_1$ (c) Gaussian noise $y_2 = y + \zeta_2$

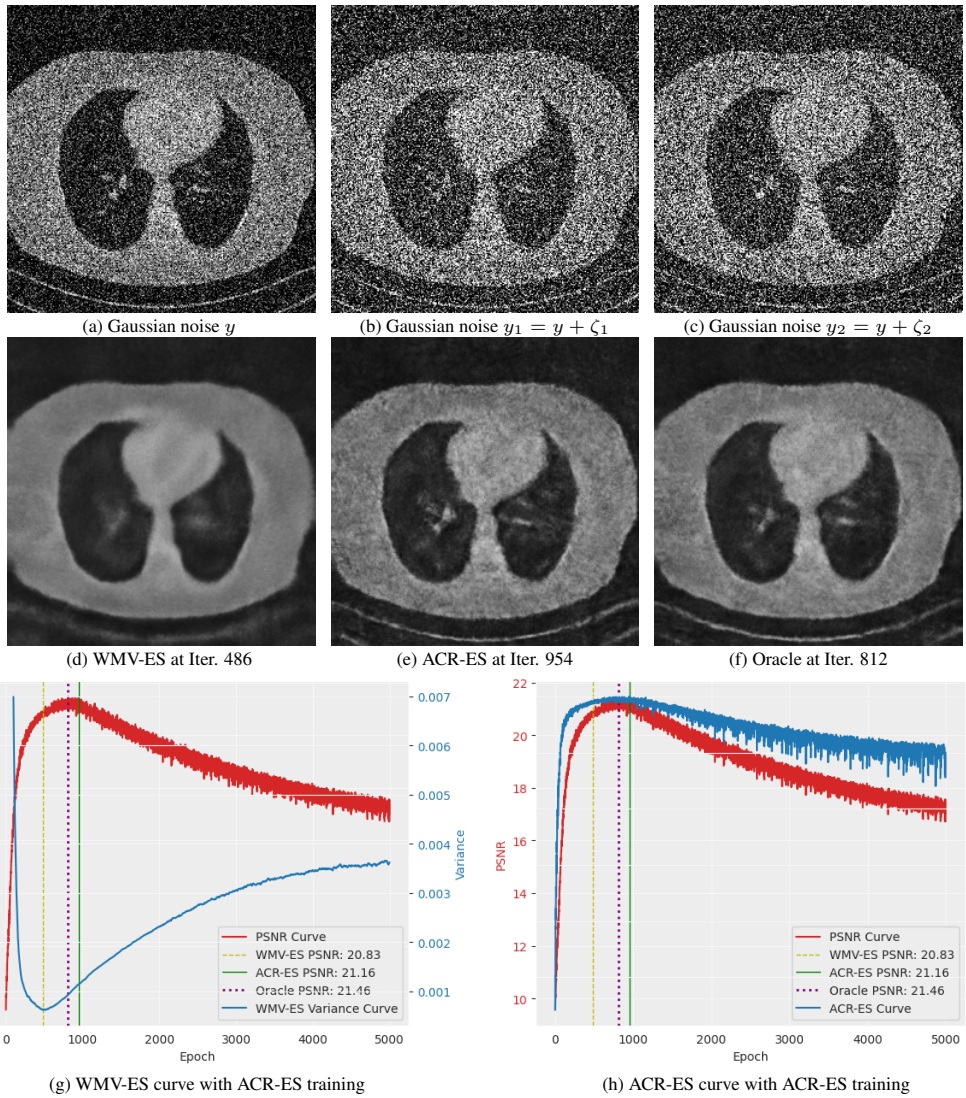
(d) WMV-ES at Iter. 534 (e) ACR-ES at Iter. 1004 (f) Oracle at Iter. 988

(g) WMV-ES curve with ACR-ES training (h) ACR-ES curve with ACR-ES training

Figure 6: Image denoising results on Gaussian noise. Comparison of WMV-ES and our ACR-ES.

Moreover, we also present the comparison of different training methods among the original DIP, the proposed ACR-ES, and the MR-ES training in Figure 9. We test a single image from the CT10 dataset under three different noise types: Gaussian, Poisson, and Impulse. Although the reference construction and input perturbations slightly modify the optimization trajectory, the overall PSNR evolution and attainable oracle performance remain comparable to those of the original DIP. Thus, the proposed methods improve early stopping mainly through more reliable stopping signals, rather than by substantially changing the DIP training process.

We further compare the proposed reference-based stopping criteria with SURE-ES and WMV-ES through representative stopping-score trajectories. Although SURE provides an unbiased estimate of the reconstruction risk under specific Gaussian-noise assumptions, its empirical minimum does not always coincide with the oracle stopping region of DIP optimization. Figure 10 shows examples on a natural image and a medical image corrupted by Gaussian noise. For the natural image, we compare the SURE curve on the original DIP trajectory with WMV-ES and MR-ES curves on the corresponding reference-based training trajectories. For the medical image, we compare SURE-ES with the stopping curve obtained under ACR-ES training. These examples show that the proposed reference-based stopping signals are better aligned with the near-oracle region than SURE-ES or WMV-ES in these representative cases. This complements the quantitative comparisons in the main



(a) Gaussian noise y (b) Gaussian noise $y_1 = y + \zeta_1$ (c) Gaussian noise $y_2 = y + \zeta_2$

(d) WMV-ES at Iter. 486 (e) ACR-ES at Iter. 954 (f) Oracle at Iter. 812

(g) WMV-ES curve with ACR-ES training (h) ACR-ES curve with ACR-ES training

Figure 7: Image denoising results on Gaussian noise. Comparison of WMV-ES and our ACR-ES.

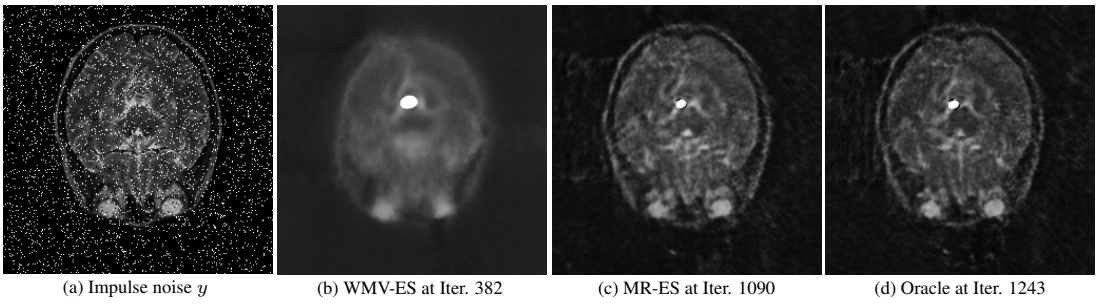
paper and further illustrates the robustness of the proposed stopping principle across image domains and training trajectories.

G.3 Full Method-specific Oracle Comparisons

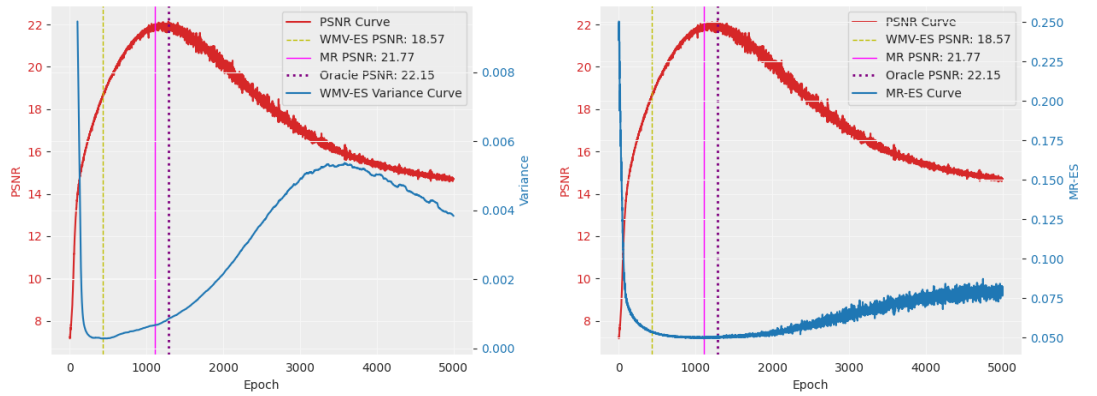
Section 4.3 summarizes the method-specific oracle analysis using group-level averages. Here, we provide the full per-image comparisons on representative natural and medical images. Additionally, we compare a recent no-reference early-stopping method cBRISQUE [30]. The purpose of this analysis is to verify that the proposed criteria reduce the PSNR Gap by selecting stopping points closer to the oracle region, rather than by lowering the attainable oracle PSNR of the corresponding trajectory.

Tables 7 and 8 report per-image results on natural and medical images, respectively. Each entry is written as ES PSNR (Oracle PSNR), where the oracle PSNR is computed from the corresponding training trajectory of each method. Thus, the selected reconstruction quality and the best attainable quality along the same method-specific trajectory can be compared simultaneously.

Table 7 reports results on selected natural images under Gaussian, Poisson, and impulse noise. Across different images and noise types, CSR-ES and ACR-ES generally achieve smaller oracle gaps than



(a) Impulse noise y (b) WMV-ES at Iter. 382 (c) MR-ES at Iter. 1090 (d) Oracle at Iter. 1243



(e) WMV-ES curve with MR-ES training (f) MR-ES curve with MR-ES training

Figure 8: Image denoising results on Impulse noise. Comparison of WMV-ES and our MR-ES.

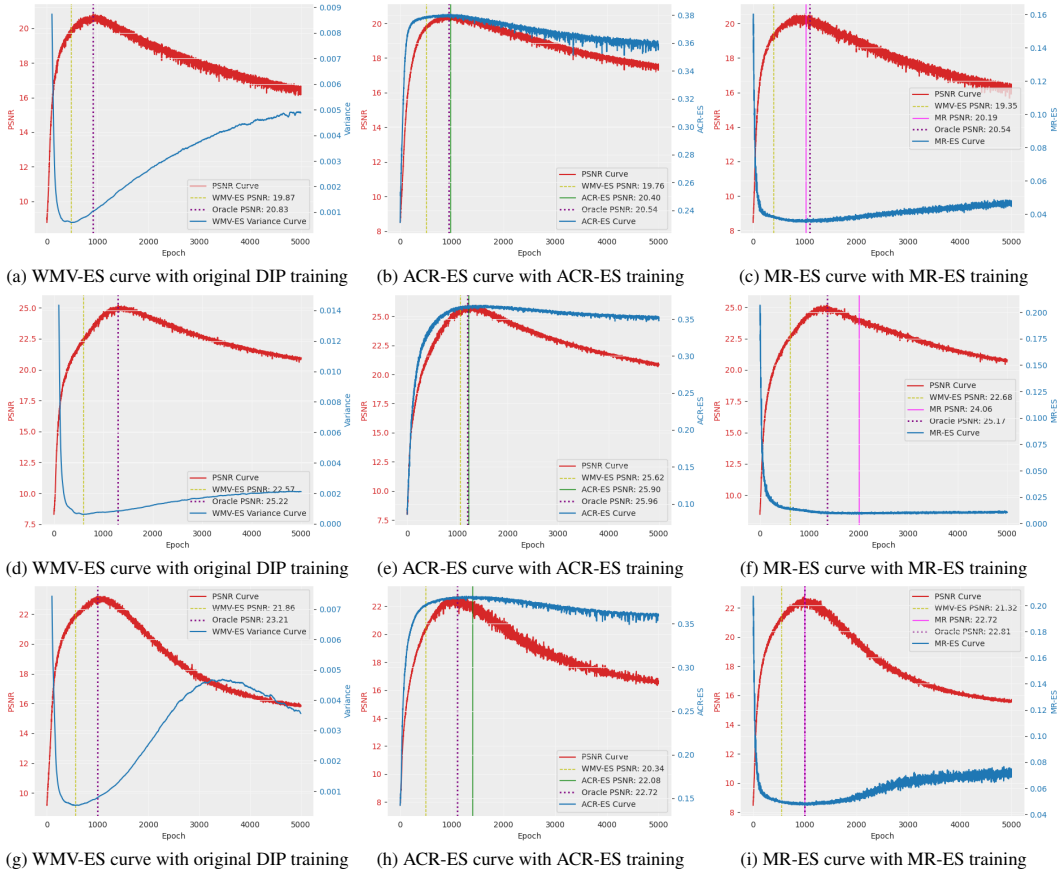
SURE-ES and WMV-ES while maintaining comparable oracle PSNR values. This indicates that the improvement is not caused by degrading the reconstruction trajectory. Instead, the proposed color-image criteria provide stopping signals that more accurately identify the near-oracle region of the corresponding trajectory.

Table 8 reports the corresponding results on selected CT and T2 images. Since CSR-ES relies on RGB channel consistency, it is not applicable to these grayscale medical images. MR-ES and ACR-ES instead provide reference-based stopping signals through mask-based and auxiliary-corruption-based references. For Gaussian and impulse noise, MR-ES and ACR-ES generally select iterations close to their own oracle trajectories. For Poisson noise, where the corruption is signal-dependent and the mask-reference signal can be less stable, ACR-ES often achieves substantially smaller gaps than SURE-ES, WMV-ES, and MR-ES. These per-image results are consistent with the averaged trends in the main paper and further support the conclusion that the proposed criteria improve stopping accuracy by better identifying the best attainable region along each trajectory.

To further strengthen the evaluation against recent early-stopping methods, we added a comparison with the recent cBRISQUE method [30]. It is a recent no-reference early-stopping method specifically designed to predict the quality of DIP checkpoints, making it the most directly comparable recent baseline to our setting. Since its code and pretrained predictors are not publicly available, we reimplemented cBRISQUE and trained its SVR/RF predictors following the paper, using 40 CBSD500 training images, disjoint from CBSD68, and 8,000 PSNR-labeled DIP checkpoints. The cBRISQUE paper evaluates only Poisson and Cauchy noise, whereas our experiments consider Gaussian, Poisson, and impulse noise. We therefore use Poisson noise, the common corruption setting, for a fair comparison. All criteria were evaluated on the same ACR-DIP trajectories, with matched architecture, Poisson noise (peak=10), optimizer, learning rate, and iterations. The results are presented in Table 9.

To assess the practical applicability of the proposed stopping criteria, we evaluate their computational overhead and reconstruction quality under multiple image-quality metrics. In addition to PSNR, we report MS-SSIM and LPIPS to examine whether the selected stopping points also yield favorable structural and perceptual quality in Table 10.

We conduct additional Gaussian denoising experiments on 10 CT images with noise level $\sigma = 0.26$, using the same budget of 5,000 iterations for all methods. To evaluate computational efficiency, each



(a) WMV-ES curve with original DIP training (b) ACR-ES curve with ACR-ES training (c) MR-ES curve with MR-ES training

(d) WMV-ES curve with original DIP training (e) ACR-ES curve with ACR-ES training (f) MR-ES curve with MR-ES training

(g) WMV-ES curve with original DIP training (h) ACR-ES curve with ACR-ES training (i) MR-ES curve with MR-ES training

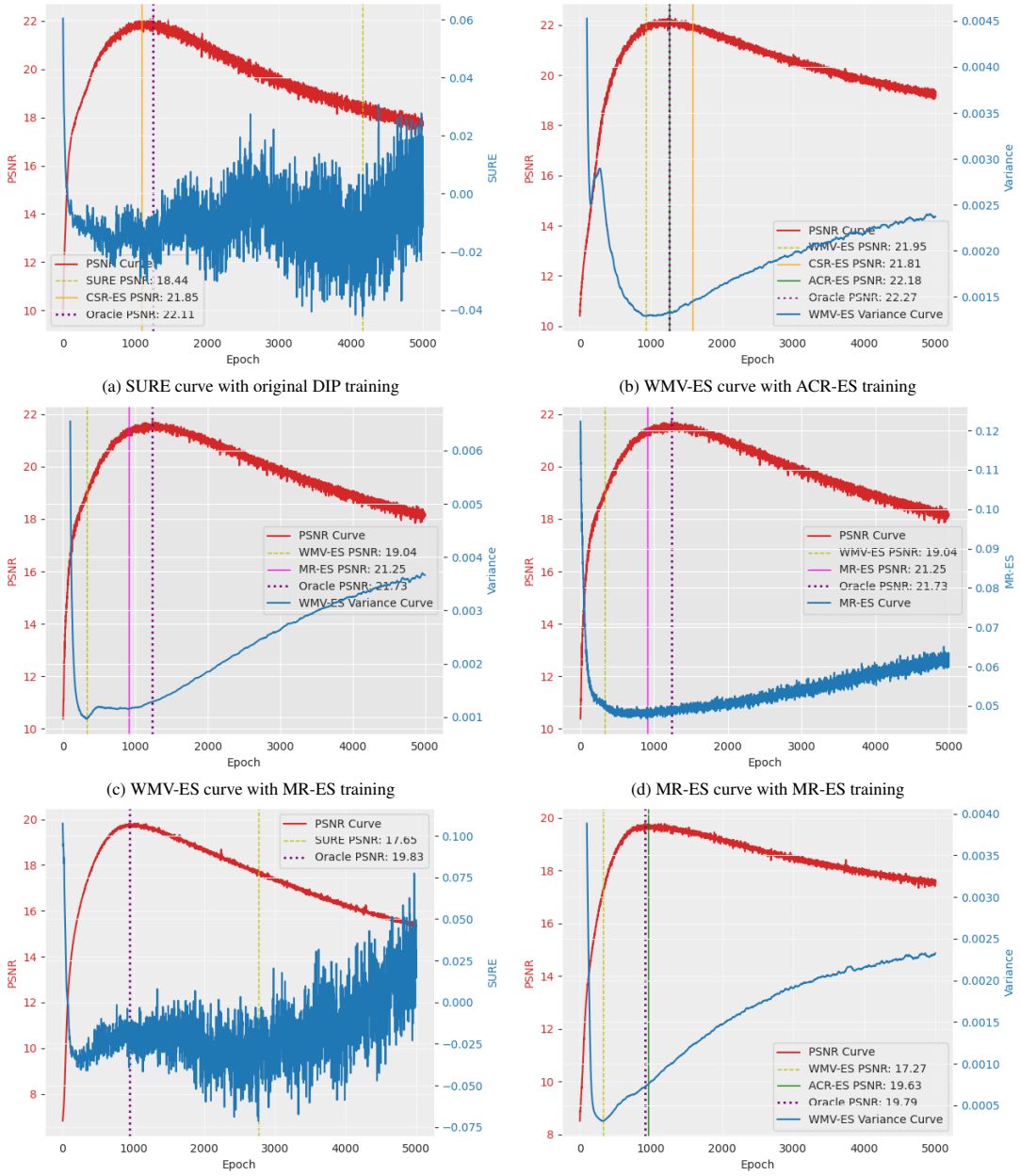
Figure 9: Comparison of different training curves of WMV-ES and the proposed reference-based stopping criteria. Test on CT01 image corrupted by Gaussian noise in the first row; Poisson noise in the second row; and Impulse noise in the third row.

method is run three times on each image. Runtime is first averaged across the three runs for each image and then reported as the mean ± standard deviation across the 10 images. Ground-truth-based metric evaluations and oracle computations are excluded from the runtime measurements. Reconstruction quality is evaluated in a separate fourth run and reported as the mean ± standard deviation across the 10 images. For comparison, we also report MS-SSIM and LPIPS at the PSNR-oracle iteration, defined as the iteration achieving the highest PSNR along the corresponding reconstruction trajectory.

As shown in Table 10, MR-ES and ACR-ES introduce only 0.9% and 1.8% additional runtime relative to WMV-ES, respectively. Both methods improve stopping accuracy in terms of PSNR, reducing the average oracle gap from 0.80 dB for WMV-ES to 0.26 dB and 0.34 dB, respectively. MR-ES achieves the highest average PSNR of 20.42 dB, while ACR-ES achieves the highest MS-SSIM (0.8019) and lowest LPIPS (0.4523) among the evaluated stopping methods, while also improving PSNR over WMV-ES. Notably, the MS-SSIM and LPIPS values at the PSNR-oracle iterations are not necessarily optimal under their respective metrics, highlighting the metric-dependent nature of reconstruction quality. Overall, these results demonstrate that the proposed stopping criteria improve PSNR-based stopping accuracy with negligible computational overhead, while ACR-ES also exhibits favorable performance across the additional quality metrics.

G.4 Ablation Study: MSE-based and Bias-based Reference Scores

We next study the effect of different reference scores. The proposed framework uses different reference constructions for natural color images and grayscale medical images. For natural color images, ACR-ES uses a bias-based cross-reference score. For grayscale medical images, MR-ES uses an MSE-based mask-reference score. This section justifies these choices.



(a) SURE curve with original DIP training

(b) WMV-ES curve with ACR-ES training

(c) WMV-ES curve with MR-ES training

(d) MR-ES curve with MR-ES training

(e) SURE curve with original DIP training

(f) WMS-ES curve with ACR training

Figure 10: Comparison of different training curves of SURE, WMV-ES, and the proposed reference-based stopping criteria. First and second row are the nature image from Set3 corrupted by Gaussian noise, the third row is the CT image from T2-20 corrupted by Gaussian noise.

Table 7: Representative early-stopping comparison on selected natural images. Each PSNR entry is reported as ES PSNR (Oracle PSNR), where the oracle PSNR is computed from the corresponding training trajectory of each method. Gap denotes Oracle PSNR minus ES PSNR; a smaller Gap indicates a more accurate stopping point.

Noise	Image	SURE-ES		WMV-ES		MR-ES		CSR-ES		ACR-ES	
		PSNR (Oracle)	Gap ↓	PSNR (Oracle)	Gap ↓	PSNR (Oracle)	Gap ↓	PSNR (Oracle)	Gap ↓	PSNR (Oracle)	Gap ↓
Gaussian	Leaves	20.06 (21.43)	1.37	20.64 (20.89)	0.25	20.62 (20.84)	0.22	20.59 (20.70)	0.11	**20.62 (20.70)**	**0.08**
	Starfish	21.31 (22.04)	0.73	18.96 (21.73)	2.77	21.23 (21.72)	0.49	21.39 (21.69)	0.30	**21.52 (21.69)**	**0.17**
	Baboon	19.50 (19.92)	0.42	18.30 (19.63)	1.33	19.37 (19.62)	0.25	19.22 (19.50)	0.28	**19.42 (19.50)**	**0.08**
	Barbara	21.81 (23.01)	1.20	21.61 (22.66)	1.05	22.30 (22.58)	0.29	**22.46 (22.55)**	**0.09**	22.21 (22.55)	0.34
	Coastguard	22.82 (23.30)	0.48	21.87 (23.10)	1.22	21.95 (22.75)	0.80	**22.77 (22.86)**	**0.09**	22.59 (22.86)	0.27
	Comic	18.46 (20.58)	2.11	17.94 (20.07)	2.12	19.76 (19.86)	0.11	19.81 (19.94)	0.13	**19.91 (19.94)**	**0.03**
	Face	22.03 (22.72)	0.69	22.39 (22.66)	0.27	22.26 (22.49)	0.23	22.27 (22.51)	0.24	**22.42 (22.51)**	**0.09**
	Flowers	21.93 (22.35)	0.42	21.07 (21.90)	0.83	21.39 (21.72)	0.34	**21.62 (21.82)**	**0.21**	21.61 (21.82)	0.21
	Foreman	20.36 (24.06)	3.71	23.50 (23.88)	0.38	23.47 (23.79)	0.31	**23.71 (23.82)**	**0.12**	23.63 (23.82)	0.19
	Zebra	22.59 (23.04)	0.45	22.09 (22.61)	0.52	22.28 (22.49)	0.22	**22.33 (22.45)**	**0.12**	22.13 (22.45)	0.32
Poisson	Leaves	20.39 (20.83)	0.44	20.33 (20.47)	0.14	20.32 (20.43)	0.11	**20.92 (21.03)**	**0.11**	20.89 (21.03)	0.13
	Starfish	22.92 (23.16)	0.24	22.15 (22.69)	0.54	22.69 (22.86)	0.18	23.40 (23.51)	0.11	**23.42 (23.51)**	**0.09**
	Baboon	18.90 (20.97)	2.07	18.66 (20.56)	1.90	20.37 (20.54)	0.17	**20.46 (20.54)**	**0.08**	20.41 (20.54)	0.13
	Barbara	23.92 (24.02)	0.10	23.29 (23.56)	0.27	23.21 (23.48)	0.27	24.75 (24.85)	0.11	24.78 (24.85)	0.07
	Coastguard	24.04 (24.47)	0.43	22.93 (23.83)	0.91	23.58 (23.75)	0.17	**24.62 (24.75)**	**0.13**	24.61 (24.75)	0.14
	Comic	19.04 (20.90)	1.87	17.75 (20.35)	2.59	19.84 (20.39)	0.55	21.38 (21.52)	0.14	**21.40 (21.52)**	**0.12**
	Face	26.30 (26.49)	0.19	25.50 (26.17)	0.67	25.89 (26.27)	0.38	26.59 (26.79)	0.20	**26.73 (26.79)**	**0.06**
	Flowers	21.71 (23.91)	2.20	22.89 (23.55)	0.65	22.99 (23.33)	0.34	**24.47 (24.54)**	**0.07**	24.39 (24.54)	0.14
	Foreman	23.27 (23.62)	0.34	23.15 (23.31)	0.15	23.07 (23.25)	0.18	23.69 (23.84)	0.15	**23.70 (23.84)**	**0.14**
	Zebra	22.53 (24.68)	2.14	23.95 (24.11)	0.17	**24.25 (24.31)**	**0.06**	24.90 (25.24)	0.33	25.01 (25.24)	0.22
Impulse	Leaves	17.79 (23.19)	5.40	22.18 (22.34)	0.16	21.59 (22.25)	0.66	**22.12 (22.17)**	**0.04**	21.88 (22.17)	0.28
	Starfish	19.10 (23.92)	4.82	22.68 (23.32)	0.64	22.99 (23.11)	0.13	23.04 (23.38)	0.34	**23.12 (23.38)**	**0.26**
	Baboon	20.74 (21.00)	0.27	18.53 (20.50)	1.97	20.03 (20.31)	0.28	**20.34 (20.36)**	**0.02**	20.30 (20.36)	0.06
	Barbara	23.69 (24.17)	0.49	23.32 (23.90)	0.58	23.34 (23.69)	0.31	**23.49 (23.65)**	**0.16**	23.32 (23.65)	0.33
	Coastguard	20.56 (24.65)	4.09	23.05 (24.26)	1.22	23.95 (24.18)	0.23	**24.12 (24.28)**	**0.15**	23.68 (24.28)	0.60
	Comic	19.86 (22.15)	2.30	18.14 (21.45)	3.31	21.00 (21.17)	0.17	**20.98 (21.13)**	**0.16**	20.93 (21.13)	0.20
	Face	16.85 (24.86)	8.01	24.27 (24.81)	0.54	24.18 (24.75)	0.57	**24.59 (24.75)**	**0.17**	23.73 (24.75)	1.02
	Flowers	21.87 (23.78)	1.91	22.63 (23.23)	0.60	22.90 (23.10)	0.20	**22.98 (23.11)**	**0.13**	22.75 (23.11)	0.36
	Foreman	20.17 (26.36)	6.19	25.68 (26.03)	0.35	25.52 (25.90)	0.38	**25.96 (26.07)**	**0.11**	25.41 (26.07)	0.66
	Zebra	22.86 (24.99)	2.13	23.98 (24.34)	0.36	23.90 (24.14)	0.24	**24.00 (24.24)**	**0.24**	23.79 (24.24)	0.44

For natural color images, we compare bias-based and MSE-based variants of ACR-ES. Table 11 shows that the bias-based ACR-ES consistently reduces the PSNR Gap compared with the MSE-based variant, especially under Gaussian and impulse noise. This suggests that, for RGB images, cross-reference bias information captures useful stopping behavior beyond direct MSE consistency.

For grayscale medical images, we compare MSE-based and bias-based variants of MR-ES. Table 12 shows that the MSE-based mask-reference criterion is generally more reliable than the bias-based variant. This is expected because single-channel medical images do not provide the same multi-channel structure as RGB images, making direct mask-reference consistency a more stable validation signal.

Together, these ablations justify the design used in the main paper: bias-based ACR-ES for natural color images and MSE-based MR-ES for grayscale medical images.

G.5 Sensitivity to Auxiliary Noise Levels

ACR-ES constructs auxiliary corrupted observations and therefore introduces auxiliary noise levels. A natural question is whether the method requires precise tuning of these levels. In the main experiments, we use the default noise level setting

$$\tau_1 = \tau_2 = 1.25\sigma,$$

where σ denotes the original noise level. Here, we study the sensitivity of ACR-ES to both auxiliary-noise magnitude and mismatch.

Table 8: Representative early-stopping comparison on selected medical images. Each PSNR entry is reported as ES PSNR (Oracle PSNR), where the oracle PSNR is computed from the corresponding training trajectory of each method. Gap denotes Oracle PSNR minus ES PSNR; a smaller Gap indicates a more accurate stopping point.

Noise	Image	SURE-ES		WMV-ES		MR-ES		ACR-ES	
		PSNR (Oracle)	Gap ↓	PSNR (Oracle)	Gap ↓	PSNR (Oracle)	Gap ↓	PSNR (Oracle)	Gap ↓
Gaussian	CT01	19.34 (20.92)	1.59	19.46 (20.49)	1.03	**20.35 (20.47)**	**0.12**	20.52 (20.68)	0.16
	CT05	20.42 (20.79)	0.37	19.93 (20.95)	1.02	20.47 (20.67)	0.20	**20.49 (20.63)**	**0.14**
	CT06	18.04 (20.50)	2.46	19.95 (20.64)	0.69	19.78 (20.25)	0.47	**20.22 (20.45)**	**0.23**
	CT08	15.26 (21.64)	6.38	21.12 (21.41)	0.28	20.12 (21.17)	1.05	**21.11 (21.30)**	**0.20**
	CT09	16.39 (20.55)	4.17	20.16 (20.57)	0.40	19.85 (20.21)	0.36	**19.96 (20.20)**	**0.24**
	T2-23	19.58 (19.76)	0.19	16.77 (19.57)	2.80	19.23 (19.64)	0.41	**19.37 (19.50)**	**0.13**
	T2-633	15.31 (19.93)	4.62	15.87 (19.94)	4.07	**19.72 (19.76)**	**0.04**	19.58 (19.75)	0.17
	T2-1173	15.27 (19.88)	4.61	17.17 (19.87)	2.70	19.68 (19.85)	0.17	**19.63 (19.79)**	**0.17**
	T2-1783	15.60 (19.86)	4.26	16.72 (19.72)	3.00	19.23 (19.53)	0.30	**19.40 (19.58)**	**0.18**
	T2-2273	18.52 (20.06)	1.54	16.72 (19.95)	3.23	19.71 (19.79)	0.08	**19.86 (19.94)**	**0.08**
Poisson	CT01	19.79 (25.66)	5.87	23.84 (26.02)	2.18	24.05 (25.05)	1.01	**26.07 (26.46)**	**0.40**
	CT05	22.71 (27.15)	4.44	24.17 (28.03)	3.86	24.63 (27.12)	2.48	**29.26 (29.60)**	**0.35**
	CT06	20.88 (26.51)	5.63	22.36 (26.11)	3.75	**25.30 (26.00)**	**0.70**	25.94 (27.18)	1.24
	CT08	22.26 (26.25)	3.99	24.27 (25.94)	1.67	23.99 (25.58)	1.59	**26.28 (26.39)**	**0.10**
	CT09	18.82 (26.31)	7.49	22.76 (26.34)	3.57	23.86 (26.11)	2.25	**26.86 (27.63)**	**0.77**
	T2-23	24.44 (25.97)	1.53	18.55 (25.96)	7.41	23.35 (25.87)	2.52	**25.60 (26.92)**	**1.32**
	T2-633	17.41 (25.60)	8.18	17.95 (26.55)	8.60	23.14 (26.48)	3.34	**27.11 (28.14)**	**1.04**
	T2-1173	24.59 (25.99)	1.39	20.30 (27.37)	7.07	23.48 (26.66)	3.17	**27.29 (28.25)**	**0.96**
	T2-1783	23.37 (24.80)	1.43	17.91 (25.62)	7.71	23.00 (25.45)	2.45	**25.12 (26.36)**	**1.24**
	T2-2273	22.56 (25.98)	3.42	17.94 (25.91)	7.97	22.86 (25.74)	2.88	**25.94 (26.95)**	**1.01**
Impulse	CT01	14.95 (22.98)	8.03	21.23 (22.90)	1.66	22.38 (22.74)	0.36	**22.14 (22.38)**	**0.24**
	CT05	14.50 (22.92)	8.42	23.04 (23.95)	0.92	**22.81 (23.06)**	**0.26**	22.45 (23.04)	0.59
	CT06	14.39 (23.26)	8.87	21.15 (23.05)	1.90	22.09 (22.83)	0.74	**22.02 (22.35)**	**0.32**
	CT08	15.71 (24.26)	8.55	23.32 (23.85)	0.53	23.84 (24.10)	0.25	**23.43 (23.67)**	**0.24**
	CT09	14.63 (23.24)	8.62	21.00 (22.85)	1.86	22.56 (22.86)	0.30	**22.49 (22.75)**	**0.26**
	T2-23	14.64 (21.99)	7.35	18.15 (22.01)	3.86	21.40 (21.88)	0.48	**21.55 (21.67)**	**0.12**
	T2-633	14.51 (22.73)	8.22	21.90 (22.53)	0.63	**22.19 (22.31)**	**0.12**	22.14 (22.26)	0.13
	T2-3	14.18 (22.60)	8.42	19.81 (22.97)	3.16	**22.10 (22.34)**	**0.24**	21.30 (21.96)	0.67
	T2-1783	14.15 (21.98)	7.83	18.02 (22.02)	4.00	21.51 (21.80)	0.29	**21.21 (21.45)**	**0.24**
	T2-2273	14.20 (22.30)	8.10	19.18 (22.21)	3.03	**21.86 (22.07)**	**0.21**	20.48 (21.92)	1.43

Table 13 studies the effect of varying the auxiliary noise magnitude by setting $\tau_1 = \tau_2 = 1.25\sigma + c$ under Gaussian, Poisson, and impulse noise. The PSNR and PSNR Gap remain stable across different values of c, showing that ACR-ES is not overly sensitive to the precise auxiliary-noise strength.

We further evaluate mismatched auxiliary noise levels in Table 14, where τ_1 and τ_2 are varied independently. The diagonal entries correspond to matched auxiliary noise levels, and the highlighted entry corresponds to the default setting $\tau_1 = \tau_2 = 1.25\sigma$. The off-diagonal entries show that ACR-ES remains effective even when $\tau_1 \neq \tau_2$.

Moreover, we test whether ACR-ES requires the auxiliary corrupted references to have the same noise type as the observed image. In this study, the observed image y is fixed as the same Gaussian-corrupted input, while the auxiliary references y_1 and y_2 are generated using Gaussian, Poisson, or impulse noise. Since the auxiliary references participate in the ACR-ES construction, changing their noise type may induce a different optimization trajectory; therefore, the oracle is computed separately for each row. As shown in Table 15, Poisson auxiliary references remain effective and achieve a gap comparable to the matched Gaussian setting on average. Impulse auxiliary references are less stable, mainly due to failures on Starfish and Butterfly, but still work well on Leaves. These results suggest that ACR-ES is reasonably robust to moderate auxiliary noise-type mismatch, although impulse-type auxiliary corruptions can be less reliable.

Table 9: Comparison of CSR-ES and ACR-ES with cBRISQUE and WMV-ES on natural-image datasets under Poisson noise (peak $= 10$). Values are mean $\pm$ standard deviation across images. Oracle uses ground truth and is shown only as a reference. Bold denotes the best non-oracle result.

Dataset	Method	PSNR ↑	PSNR gap ↓	SSIM ↑
CBSD68	WMV-ES	23.10 ± 3.32	1.69 ± 1.80	0.581 ± 0.183
	cBRISQUE†	22.35 ± 3.46	2.44 ± 2.39	0.552 ± 0.165
	CSR-ES (Ours)	24.18 ± 2.31	0.61 ± 1.03	0.678 ± 0.081
	ACR-ES (Ours)	$\mathbf{24.66 \pm 2.33}$	$\mathbf{0.13 \pm 0.07}$	$\mathbf{0.694 \pm 0.077}$
	Oracle	24.79 ± 2.34	0.00 ± 0.00	0.697 ± 0.076
Set18	WMV-ES	25.64 ± 3.48	1.06 ± 2.06	0.736 ± 0.090
	cBRISQUE†	21.29 ± 4.73	5.41 ± 5.66	0.574 ± 0.179
	CSR-ES (Ours)	25.37 ± 2.54	1.33 ± 1.82	0.714 ± 0.110
	ACR-ES (Ours)	$\mathbf{26.55 \pm 2.32}$	$\mathbf{0.15 \pm 0.07}$	$\mathbf{0.763 \pm 0.043}$
	Oracle	26.70 ± 2.30	0.00 ± 0.00	0.766 ± 0.043
Set3c	WMV-ES	22.93 ± 1.79	0.21 ± 0.08	0.784 ± 0.063
	cBRISQUE†	21.98 ± 1.67	1.16 ± 0.92	0.771 ± 0.052
	CSR-ES (Ours)	22.68 ± 1.70	0.46 ± 0.36	0.776 ± 0.061
	ACR-ES (Ours)	$\mathbf{22.98 \pm 1.83}$	$\mathbf{0.15 \pm 0.04}$	$\mathbf{0.788 \pm 0.059}$
	Oracle	23.14 ± 1.87	0.00 ± 0.00	0.792 ± 0.066

† cBRISQUE uses offline supervised training with clean-image/ground-truth PSNR labels. The other stopping rules do not require a pretrained predictor.

Table 10: Average Gaussian denoising results on CT10 at $\sigma = 0.26$ with a common 5,000-iteration budget. Values are mean $\pm$ standard deviation across 10 CT images. Runtime is the mean of three runs per image, summarized across images; ground-truth computations are excluded from timing. "At PSNR oracle" evaluates each metric at the iterate that maximizes PSNR, not at an independently selected metric-specific oracle. Bold denotes the best selected-stop metric or smallest PSNR gap among stopping methods.

			At selected stop			At PSNR oracle			
Method	Runtime (s/image)	Overhead vs. WMV-ES	PSNR ↑	MS-SSIM ↑	LPIPS ↓	PSNR ↑	MS-SSIM ↑	LPIPS ↓	PSNR gap ↓
WMV-ES	54.62 ±1.37	–	19.91 ±0.59	0.7983 ±0.0204	0.4555 ±0.0624	20.72 ±0.47	0.8072 ±0.0180	0.4419 ±0.0624	0.80 ±0.31
MR-ES (Ours)	55.14 ±1.35	+0.9%	**20.42 ±0.44**	0.7881 ±0.0275	0.4614 ±0.0676	20.68 ±0.48	0.8072 ±0.0266	0.4469 ±0.0589	**0.26 ±0.11**
ACR-ES (Ours)	55.62 ±1.04	+1.8%	20.21 ±0.62	**0.8019 ±0.0167**	**0.4523 ±0.0524**	20.55 ±0.49	0.7966 ±0.0239	0.4596 ±0.0614	0.34 ±0.24

Table 16 studies a reversed noise-type mismatch setting, where the auxiliary references y_1 and y_2 are fixed to Gaussian-corrupted observations, while the observed image y is corrupted by Gaussian, Poisson, or impulse noise. The Gaussian case is the matched setting, and the Poisson and impulse cases test whether ACR-ES remains effective when the observation and auxiliary references have different noise types. ACR-ES remains reasonably robust under this mismatch: the average gap increases from 0.33 dB in the matched Gaussian case to 0.49 dB for Poisson and 0.61 dB for impulse. This indicates that ACR-ES does not require an exact match between the noise type of y and that of the auxiliary references, although matched Gaussian references yield the most stable stopping behavior.

These results indicate that ACR-ES is robust to moderate auxiliary-noise scaling and mismatch, and does not require precise tuning of the auxiliary corruption levels.

G.6 Early stopping beyond vanilla DIP

It is important to evaluate whether the proposed early-stopping method generalizes beyond vanilla DIP and to examine how explicit regularization affects its stopping behavior. We therefore added

Table 11: Comparison between bias-based and MSE-based ACR-ES on natural image datasets. PSNR and PSNR Gap are reported as mean ± standard deviation across images. Positive Δ Gap indicates that ACR-ES (Bias) is closer to the oracle stopping point.

Noise	Dataset	ACR-ES (Bias)		ACR-ES (MSE)		Δ Gap ↑
		PSNR ↑	Gap ↓	PSNR ↑	Gap ↓	
Gaussian	Set14	**22.26(±1.71)**	**0.25(±0.13)**	21.44(±1.67)	1.06(±0.79)	0.81
	Set18	**22.33(±1.52)**	**0.21(±0.12)**	21.65(±0.84)	0.89(±1.24)	0.68
	CBSD68	**22.33(±1.99)**	**0.30(±0.28)**	21.15(±1.58)	1.48(±1.44)	1.18
	Urban100	**21.36(±1.31)**	**0.20(±0.10)**	20.60(±1.01)	0.97(±0.65)	0.76
Poisson	Set14	**24.52(±2.35)**	**0.11(±0.05)**	24.52(±2.35)	0.11(±0.07)	0.00
	Set18	**26.55(±2.33)**	**0.15(±0.10)**	26.53(±2.32)	0.17(±0.11)	0.02
	CBSD68	**24.66(±2.33)**	**0.14(±0.09)**	24.66(±2.32)	0.14(±0.08)	0.00
	Urban100	24.08(±1.66)	0.18(±0.12)	**24.09(±1.64)**	**0.17(±0.10)**	-0.01
Impulse	Set14	**23.65(±1.74)**	**0.49(±0.23)**	22.50(±1.77)	1.64(±1.56)	1.15
	Set18	**24.09(±1.52)**	**0.53(±0.21)**	23.11(±0.97)	1.50(±0.76)	0.97
	CBSD68	**23.63(±2.02)**	**0.44(±0.25)**	22.29(±1.36)	1.78(±1.08)	1.34
	Urban100	**22.86(±1.42)**	**0.34(±0.17)**	22.23(±1.08)	0.97(±0.66)	0.63

Table 12: Comparison between the MSE-based and bias-based mask-reference early stopping criteria on medical datasets. PSNR and PSNR Gap are reported as mean ± standard deviation across images. Positive Δ Gap=Gap(Bias)-Gap(MSE) indicates that the MSE-based criterion is closer to the oracle stopping point.

Noise	Dataset	MR-ES (MSE)		MR-ES (Bias)		Δ Gap ↑
		PSNR ↑	Gap ↓	PSNR ↑	Gap ↓	
Gaussian	MRI15	**21.60(±0.69)**	**0.41(±0.29)**	19.69(±1.69)	2.33(±1.74)	1.92
	CT10	**20.22(±0.47)**	**0.35(±0.28)**	19.66(±1.36)	0.92(±1.26)	0.57
	PD20	**19.94(±0.36)**	**0.33(±0.24)**	18.03(±0.92)	2.24(±0.93)	1.91
	T2-20	**19.62(±0.19)**	**0.16(±0.09)**	18.33(±1.01)	1.45(±0.98)	1.29
Poisson	MRI15	**24.78(±1.27)**	**1.35(±1.04)**	23.33(±1.83)	2.80(±1.65)	1.45
	CT10	**24.56(±1.30)**	**2.03(±0.85)**	23.36(±1.71)	3.24(±1.60)	1.21
	PD20	**21.15(±0.37)**	4.49(±0.37)	21.29(±0.47)	**4.34(±0.36)**	-0.15
	T2-20	23.26(±0.47)	3.05(±0.36)	**23.53(±0.78)**	**2.78(±0.72)**	-0.27
Impulse	MRI15	**23.49(±0.52)**	**0.46(±0.47)**	18.98(±2.91)	4.96(±2.96)	4.50
	CT10	**22.82(±0.70)**	**0.31(±0.17)**	16.77(±1.81)	6.36(±1.50)	6.05
	PD20	**23.12(±0.28)**	**0.26(±0.19)**	16.51(±1.28)	6.87(±1.34)	6.61
	T2-20	**21.89(±0.26)**	**0.31(±0.16)**	16.72(±1.95)	5.48(±1.93)	5.17

experiments with two substantially different DIP variants: (i) aSeqDIP [24], an input-adaptive DIP method with sequential input updates and an explicit autoencoding regularizer, and (ii) Deep Decoder [31], an architecturally underparameterized reconstruction network. For aSeqDIP, we compare the regularized setting $\lambda = 1$ with an ablated setting $\lambda = 0$ that removes the explicit regularization term while retaining the remaining aSeqDIP optimization procedure.

At the moderate Gaussian noise level ($\sigma = 15/255$, the original setting in [24]), Table 17 shows that regularization reduces late-stage degradation and shifts the oracle peak to later iterations. Without regularization, the average PSNR decreases from an oracle value of 31.39 dB to 30.21 dB at 10k iterations, corresponding to a 1.18 dB degradation. With regularization, this degradation is reduced to 0.56 dB, from 31.12 dB to 30.56 dB.

Table 13: Sensitivity of ACR-ES on Set3 to the additional noise level c. Results are reported as mean $\pm$ standard deviation across images. PSNR Gap denotes the difference between Oracle PSNR and ACR-ES PSNR. For Gaussian noise, the level of noise in the observation y is $\sigma = 0.26$ and the level of noise added to the augmented channels is $\tau = 1.25\sigma + 0.1c$. For Poisson noise, the noise in y follows $\mathcal{P}(\lambda)$ with $\lambda = 10$ and the noise added to augmented channel follows $\mathcal{P}(\tau)$ with $\tau = 1.25\lambda + c$. For impulse noise, the probability of measurement corruption is $p = 0.1$ in the observation y and $\tau = 1.25p + 0.1c$ in the augmented channel.

Noise	c	ACR-ES PSNR ↑	Oracle PSNR ↑	Gap ↓
Gaussian	0	21.44(±0.78)	21.55(±0.80)	**0.11(±0.05)**
	0.10	21.42(±0.84)	21.56(±0.77)	0.14(±0.08)
	0.15	21.34(±0.85)	21.54(±0.80)	0.20(±0.06)
	0.20	**21.44(±0.87)**	21.55(±0.86)	**0.11(±0.08)**
Poisson	0	22.91(±1.82)	23.05(±1.84)	0.14(±0.06)
	0.10	**23.00(±1.80)**	23.06(±1.74)	**0.06(±0.06)**
	0.15	22.81(±1.73)	22.97(±1.74)	0.17(±0.09)
	0.20	22.89(±1.66)	23.07(±1.69)	0.18(±0.18)
Impulse	0	22.72(±0.73)	23.05(±0.77)	0.33(±0.10)
	0.10	**22.84(±0.62)**	22.99(±0.63)	**0.14(±0.08)**
	0.15	22.81(±0.65)	22.97(±0.60)	0.16(±0.06)
	0.20	22.73(±0.62)	22.95(±0.63)	0.22(±0.01)

Table 14: Sensitivity of ACR-ES on Set3 to auxiliary-noise mismatch under Gaussian noise. Each entry is reported as ES PSNR (Oracle PSNR). The default setting used throughout the paper is $\tau_1 = \tau_2 = 1.25\sigma$, highlighted in purple. Off-diagonal entries correspond to mismatched auxiliary noise levels $\tau_1 \neq \tau_2$. ACR-ES uses the bias score $\mathrm{bias}(y_2, \hat{x}_1)$. Gray cells denote matched but non-default settings $\tau_1 = \tau_2$.

τ_2 \ τ_1	1.0σ	1.25σ	1.5σ
1.0σ	20.73 (21.83)	21.46 (21.60)	**21.65 (21.65)**
1.25σ	21.19 (21.69)	**21.52 (21.69)**	**21.52 (21.69)**
1.5σ	20.73 (21.68)	**21.34 (21.68)**	21.32 (21.65)

Table 17 indicates that CSR-ES remains beneficial in both settings. In addition, the unregularized model combined with CSR-ES achieves 30.97 dB, outperforming the final regularized reconstruction of 30.56 dB by 0.41 dB. This result indicates that accurate stopping can be competitive with direct overfitting mitigation.

We further evaluated CSR-ES with aSeqDIP under a heavier Gaussian-noise setting ($\sigma = 0.26$, the setting in our paper). The results show that aSeqDIP still exhibits substantial late-stage overfitting. After 10k iterations, the average PSNR drop from the oracle peak is 5.19 dB without regularization and 3.48 dB with regularization. As shown in Table 18, the average oracle-to-final degradation is 5.19 dB without regularization and 3.48 dB with regularization. Thus, the explicit regularizer mitigates but does not eliminate overfitting.

We additionally evaluate Deep Decoder [31], which controls overfitting through architectural underparameterization rather than an explicit loss regularizer. Table 19 shows that, under $\sigma = 0.26$, CSR-ES improves the final 5k reconstruction on all three Set3 images. These results demonstrate that the proposed stopping principle is applicable not only to vanilla DIP and explicitly regularized aSeqDIP, but also to an underparameterized DIP architecture.

Table 15: Auxiliary noise-type mismatch study of ACR-ES on Set3. The observed image y is fixed as the same Gaussian-corrupted input with noise level $\sigma = 0.26$. Only the auxiliary corrupted references y_1 and y_2 vary across rows: the matched Gaussian setting uses noise level 1.25×0.26, the Poisson setting uses noise level 1.25×10, and the impulse setting uses corruption probability 1.25×0.1. Since different auxiliary references induce different ACR-ES trajectories, the oracle iteration and oracle PSNR are computed separately for each row. PSNR Gap denotes the difference between the row-specific oracle PSNR and the selected PSNR.

Noise type	Image	ACR-ES			Oracle		Δ Gap
		Iter.	PSNR ↑	Gap ↓	Iter.	PSNR ↑	
Gaussian	Starfish	1026	21.50	0.18	1437	21.68	–
	Butterfly	872	22.00	0.25	1081	22.25	–
	Leaves	1101	20.39	0.31	1580	20.70	–
	Average	–	21.29($\pm$0.82)	0.25($\pm$0.07)	–	21.54($\pm$0.78)	–
Poisson	Starfish	2193	22.25	0.32	1674	22.56	0.13
	Butterfly	1712	23.60	0.12	1451	23.72	-0.13
	Leaves	2300	21.63	0.22	1949	21.85	-0.09
	Average	–	22.49($\pm$1.01)	0.22($\pm$0.10)	–	22.71($\pm$0.94)	-0.03
Impulse	Starfish	2720	20.33	1.40	1328	21.73	1.21
	Butterfly	2201	21.22	1.18	1170	22.40	0.93
	Leaves	1863	20.53	0.22	1422	20.75	-0.09
	Average	–	20.69($\pm$0.47)	0.93($\pm$0.63)	–	21.63($\pm$0.83)	0.69

H Generalization to Forward Operators

The theory in Appendix A is stated for the denoising case in which the reference and the reconstruction live in the same space. It is possible to extend the theory to a general inverse problem $y = A(x) + \xi$, with the validation curve evaluated in the measurement domain. This section records the generalization, identifies what it does and does not control, and shows that the mask-reference theorem (Theorem E.2) and the denoising theorem (Theorem A.1) are both special cases.

H.1 Setup

Let D denote the data used to fit the trajectory $\{\hat{x}_t(D)\}_{t=1}^T$. The data D may be a noisy measurement $y_1 = A(x) + \eta_1$, but more generally it can be any training input from which a DIP trajectory is produced. If the reconstruction procedure uses additional randomness (e.g. random network initialization, dropout,stochastic optimization), we fold this randomness into D or condition on it, so that $\hat{x}_t$ is a measurable function of D. Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear deterministic validation operator, and suppose an independent reference measurement

$$y_2^A \;=\; A(x) + \eta_2$$

is available, where $\eta_2 \in \mathbb{R}^m$ has independent, zero-mean, sub-Gaussian entries with variance proxy σ^2, satisfies $\mathbb{E}\eta_{2,i}^2 = \sigma^2$, and is *independent of D and hence of the fitted trajectory* $\{\hat{x}_t(D)\}$. Define the measurement-domain reconstruction error

$$R_A(t) \;:=\; \frac{1}{m}\left\|A(\hat{x}_t(D)) - A(x)\right\|^2,$$

and the operator reference loss

$$V_t^A \;:=\; \frac{1}{m}\left\|A(\hat{x}_t(D)) - y_2^A\right\|^2.$$

The image-domain risk $R(t) = \frac{1}{n}\|\hat{x}_t(D) - x\|^2$ is, in general, *not* the quantity that the reference loss tracks. We return to this point in Section H.4.

Table 16: Auxiliary noise-type mismatch study of ACR-ES on Set3 with fixed Gaussian auxiliary references. The observed image y is corrupted by different noise types: Gaussian noise with level $\sigma = 0.26$, Poisson noise with level 10, and impulse noise with corruption probability 0.1. The auxiliary corrupted references y_1 and y_2 are always generated using Gaussian noise with level 1.25×0.26. Since different observed noise types induce different DIP trajectories, the oracle iteration and oracle PSNR are computed separately for each row. PSNR Gap denotes the difference between the row-specific oracle PSNR and the selected PSNR.

Noise of y	Image	ACR-ES			Oracle		Δ Gap
		Iter.	PSNR ↑	Gap ↓	Iter.	PSNR ↑	
Gaussian (matched)	Starfish	1034	21.38	0.16	1291	21.54	–
	Butterfly	1013	22.05	0.09	1288	22.15	–
	Leaves	843	19.90	0.73	1579	20.63	–
	Average	–	21.11($\pm$1.10)	0.33($\pm$0.35)	–	21.44($\pm$0.76)	–
Poisson (mismatched)	Starfish	962	21.86	0.53	1471	22.39	0.37
	Butterfly	894	22.53	0.55	1412	23.08	0.45
	Leaves	1056	19.72	0.40	1856	20.12	-0.33
	Average	–	21.37($\pm$1.47)	0.49($\pm$0.08)	–	21.87($\pm$1.55)	0.16
Impulse (mismatched)	Starfish	965	22.45	0.64	1649	23.09	0.47
	Butterfly	886	22.76	0.63	1646	23.39	0.54
	Leaves	1073	21.44	0.56	1855	22.00	-0.17
	Average	–	22.22($\pm$0.69)	0.61($\pm$0.04)	–	22.82($\pm$0.73)	0.28

Table 17: aSeqDIP Gaussian denoising on Set3 at $\sigma = 15/255$. Final@10k denotes the reconstruction after 10,000 iterations; Oracle is the ground-truth-selected maximum PSNR and is for reference only. CSR gain is CSR minus Final@10k (dB); Oracle gap is Oracle minus CSR (dB).

Image	λ	Final@10k	CSR (Ours)	CSR Iter.	Oracle	Oracle Iter.	CSR–Final@10k Gain	Oracle–CSR Gap
Butterfly	0	29.48	**30.43**	6,519	30.76	5,499	+0.95	0.33
Leaves	0	30.55	**31.87**	6,649	32.18	6,346	+1.32	0.31
Starfish	0	**30.61**	**30.61**	9,999	31.22	6,520	+0.00	0.61
Average	–	30.21	**30.97**	–	31.39	–	+0.76	0.42
Butterfly	1	30.22	**30.37**	7,910	30.56	6,576	+0.15	0.18
Leaves	1	30.96	**31.80**	7,483	32.03	6,662	+0.84	0.23
Starfish	1	30.50	**30.54**	9,990	30.76	7,737	+0.04	0.22
Average	–	30.56	**30.91**	–	31.12	–	+0.35	0.21

H.2 Operator-form validation theorem

Theorem H.1 (Operator reference validation)**.** *Under the setup above, there exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$, uniformly over $1 \le t \le T$,*

$$\left| V_t^A - R_A(t) - \sigma^2 \right| \;\le\; C\left[M_T^A \sqrt{\frac{\log(T/\delta)}{m}} \;+\; \sigma^2 \, \frac{\log(T/\delta)}{m} \right],$$

where

$$M_T^A \;:=\; \sigma \sup_{1 \le t \le T} \sqrt{R_A(t)} \;+\; \sigma^2.$$

Proof. Condition on D and hence on the entire DIP trajectory $\{\hat{x}_t(D)\}_{t=1}^T$. For each fixed t, write

$$a_t \;:=\; A(\hat{x}_t(D)) - A(x),$$

Table 18: aSeqDIP Gaussian denoising on Set3 at the heavier noise level $\sigma = 0.26$. Final@10k denotes the reconstruction after 10,000 iterations; Oracle is the ground-truth-selected maximum PSNR and is shown for reference only. CSR gain and Oracle gap are in dB.

Image	λ	Final@10k	CSR (Ours)	CSR Iter.	Oracle	Oracle Iter.	CSR Gain	Oracle Gap
Butterfly	0	16.87	**22.10**	1,298	22.96	1,287	+5.23	0.86
Leaves	0	16.35	**21.73**	1,733	21.79	1,762	+5.38	0.06
Starfish	0	18.01	**20.86**	2,192	22.04	1,309	+2.85	1.18
Average	–	17.08	**21.57**	–	22.26	–	+4.49	0.70
Butterfly	1	19.08	**22.28**	2,080	22.53	1,601	+3.20	0.25
Leaves	1	18.66	**21.02**	3,368	21.15	3,351	+2.36	0.13
Starfish	1	17.77	**19.06**	360	22.28	1,329	+1.29	3.22
Average	–	18.50	**20.79**	–	21.99	–	+2.29	1.20

Table 19: Deep Decoder Gaussian denoising on Set3 at $\sigma = 0.26$. Final@5k denotes the reconstruction after 5,000 iterations; Oracle denotes the ground-truth-selected maximum PSNR and is for reference only. CSR gain and Oracle gap are in dB.

Image	Final@5k	CSR (Ours)	CSR Iter.	Oracle	Oracle Iter.	CSR Gain	Oracle Gap
Butterfly	19.80	**21.04**	928	21.61	950	+1.24	0.57
Leaves	18.39	**20.26**	947	20.26	947	+1.87	0.00
Starfish	19.48	**21.23**	912	21.99	436	+1.75	0.76
Average	19.22	**20.84**	–	21.28	–	+1.62	0.44

which is deterministic given D. Since $y_2^A = A(x) + \eta_2$,

$$V_t^A \;=\; \frac{1}{m}\|a_t - \eta_2\|^2 \;=\; \frac{1}{m}\|a_t\|^2 - \frac{2}{m}\langle a_t, \eta_2\rangle + \frac{1}{m}\|\eta_2\|^2,$$

so that

$$V_t^A - R_A(t) - \sigma^2 \;=\; -\frac{2}{m}\langle a_t, \eta_2\rangle \;+\; \frac{1}{m}\big(\|\eta_2\|^2 - m\sigma^2\big).$$

By hypothesis η_2 is independent of D, hence of a_t. The first term is therefore a centered sub-Gaussian linear functional in η_2 with fixed weights a_t. Thus with probability at least $1 - \delta/(2T)$,

$$\left|\frac{2}{m}\langle a_t, \eta_2\rangle\right| \;\le\; C\sigma\sqrt{R_A(t)}\sqrt{\frac{\log(T/\delta)}{m}}.$$

The second term is a centered average of sub-exponential random variables $\eta_{2,i}^2 - \sigma^2$ and is t-independent. Bernstein concentration gives, with probability at least $1 - \delta/2$,

$$\left|\frac{1}{m}\big(\|\eta_2\|^2 - m\sigma^2\big)\right| \;\le\; C\sigma^2\left[\sqrt{\frac{\log(T/\delta)}{m}} + \frac{\log(T/\delta)}{m}\right].$$

Union-bounding the first inequality over $t = 1, \ldots, T$ contributes total failure probability at most $\delta/2$; combining with the second event, both hold simultaneously with probability at least $1 - \delta$. On this event,

$$\big|V_t^A - R_A(t) - \sigma^2\big| \le C\sigma\Big(\sqrt{R_A(t)} + \sigma\Big)\sqrt{\frac{\log(T/\delta)}{m}} + C\sigma^2\frac{\log(T/\delta)}{m},$$

uniformly in t. Since $\sigma(\sqrt{R_A(t)}+\sigma) \le \sigma \sup_t \sqrt{R_A(t)}+\sigma^2 = M_T^A$, the stated bound follows. □

H.3 Oracle inequality

Corollary H.1.1 (Near-oracle stopping in the measurement domain)**.** *Under the event of Theorem H.1, define*

$$\varepsilon_m^A \;:=\; C\left[M_T^A\sqrt{\frac{\log(T/\delta)}{m}} + \sigma^2\frac{\log(T/\delta)}{m}\right].$$

Let $\hat{t} \in \arg\min_t V_t^A$ *and* $t_A^\star \in \arg\min_t R_A(t)$. *Then*

$$R_A(\hat{t}) \;\le\; R_A(t_A^\star) + 2\varepsilon_m^A.$$

Proof. On the event of Theorem H.1, $R_A(t) \le V_t^A - \sigma^2 + \varepsilon_m^A$ and $V_t^A - \sigma^2 \le R_A(t) + \varepsilon_m^A$ hold simultaneously for all t. Hence

$$R_A(\hat{t}) \le V_{\hat{t}}^A - \sigma^2 + \varepsilon_m^A \le V_{t_A^\star}^A - \sigma^2 + \varepsilon_m^A \le R_A(t_A^\star) + 2\varepsilon_m^A. \qquad \square$$

H.4 Transfer from measurement-domain risk to image-domain risk

Corollary H.1.1 controls R_A, not the image-domain risk $R(t) = \frac{1}{n}\|\hat{x}_t - x\|^2$. The two coincide when A is the identity and the dimensions match, but in general the reference loss cannot detect overfitting in directions to which A is insensitive. A transfer condition is therefore needed in the same spirit as Corollary E.2.1. We give a two-sided version that yields the cleanest comparison to the image-domain oracle, followed by a one-sided relaxation that is more honest about what is achievable for poorly conditioned A.

Corollary H.1.2 (Image-domain near-oracle stopping; two-sided transfer)**.** *Suppose that, along the trajectory,*

$$\sup_{1\le t\le T} \big|R(t) - c\,R_A(t)\big| \le \kappa$$

for some operator-dependent constants $c > 0$ *and* $\kappa \ge 0$. *Then the operator reference stopping time* $\hat{t} \in \arg\min_t V_t^A$ *satisfies*

$$R(\hat{t}) \;\le\; \min_{1\le t\le T} R(t) \;+\; 2\kappa \;+\; 2c\,\varepsilon_m^A.$$

Proof. Let $t^\star \in \arg\min_t R(t)$. By the comparison assumption, $R(\hat{t}) \le c\,R_A(\hat{t}) + \kappa$. Corollary H.1.1 gives $R_A(\hat{t}) \;\le\; R_A(t_A^\star) + 2\varepsilon_m^A \;\le\; R_A(t^\star) + 2\varepsilon_m^A$, and the comparison assumption again gives $R_A(t^\star) \le \frac{1}{c}(R(t^\star) + \kappa)$. Combining the three inequalities yields the stated bound. $\square$

Remark H.2 (One-sided relaxation)**.** *The two-sided comparison in Corollary H.1.2 is needed only for the comparison to the image-domain oracle* $R(t^\star)$. *If one is satisfied with an absolute bound on* $R(\hat{t})$ *that uses the measurement-domain oracle as the benchmark, only the one-sided condition*

$$\sup_{1\le t\le T} \big(R(t) - c\,R_A(t)\big) \le \kappa$$

is required, in which case

$$R(\hat{t}) \;\le\; c\,R_A(t_A^\star) + 2c\,\varepsilon_m^A + \kappa.$$

This one-sided form is the one that follows from a lower singular-value bound on A *alone (see Section H.5 below).*

H.5 Linear-operator interpretation

For a linear A with smallest non-zero singular value $\sigma_{\min}(A)$, decompose the reconstruction error as $\hat{x}_t - x = u_t^R + u_t^N$ with $u_t^R \in \ker(A)^\perp$ and $u_t^N \in \ker(A)$. Then $R_A(t) = \frac{1}{m}\|Au_t^R\|^2$ and $R(t) = \frac{1}{n}(\|u_t^R\|^2 + \|u_t^N\|^2)$. Using $\sigma_{\min}(A)^2\|u_t^R\|^2 \le \|Au_t^R\|^2$ gives the *one-sided* bound

$$R(t) \;\le\; \frac{m}{n\,\sigma_{\min}(A)^2}\,R_A(t) \;+\; \frac{\|u_t^N\|^2}{n}.$$

This corresponds to the one-sided comparison in Remark H.2, with $c = m/(n\sigma_{\min}(A)^2)$ and $\kappa = \sup_t \|u_t^N\|^2/n$ being controlled by whatever mechanism limits null-space artifacts along the trajectory (e.g. DIP's spectral bias, or explicit regularization).

The reverse direction of the two-sided comparison required by Corollary H.1.2 is *not* implied by $\sigma_{\min}(A)$ alone. It depends on the ratio $\sigma_{\max}(A)/\sigma_{\min}(A)$ on the active subspace, together with a bound on $\|u_t^R\|^2$ along the trajectory. In particular, the two-sided comparison holds when A is approximately isometric on the error directions traced out by the trajectory. For severely ill-conditioned A, only the weaker one-sided guarantee in Remark H.2 is available without additional structural assumptions.

H.6 Recovery of the denoising and mask-reference theorems

Theorem H.1 reduces to the denoising and mask-reference theorems for two natural choices of (D, A).

(i) **Denoising.** Take $D = y_1 = x + \eta_1$, $A = I$, $m = n$, and $y_2^A = x + \eta_2$. Then $R_A(t) = R(t)$, $V_t^A = V_t$, and Theorem H.1 reduces to Theorem A.1. The transfer condition in Corollary H.1.2 is satisfied with $c = 1$, $\kappa = 0$.

(ii) **Mask reference.** Condition on the mask M. Let $P_H : \mathbb{R}^n \to \mathbb{R}^m$, with $m = \sum_i H_i$, denote the restriction operator that extracts the held-out coordinates. Take the training data to be $D = (M, M \odot y)$, take $A = P_H$ as the validation operator, and take

$$y_2^A = P_H y = P_H x + P_H \xi.$$

The validation noise $\eta_2 = P_H \xi$ is independent of the retained training data D because the pixel noise is independent across coordinates and the mask is conditioned on (Lemma E.1). With this identification,

$$V_t^A = \frac{1}{m} \|P_H \hat{x}_t - P_H y\|^2 = \frac{1}{m} \|H \odot (\hat{x}_t - y)\|^2 = V_t^{\text{mask}},$$

so Theorem H.1 reduces to Theorem E.2, and Corollary H.1.2 reduces to the representativeness transfer in Corollary E.2.1.

The recovery of MR-ES uses the fact that the proof of Theorem H.1 does not require D to be of the form $A(x) + \eta_1$; it only requires $\eta_2 \perp D$ and the sub-Gaussian assumption on η_2. The training data (which generates D) and the validation operator A may therefore differ, as they do in MR-ES.

The unifying object is the triple (training data D, validation operator A, independent reference y_2^A with conditional mean $A(x)$ and sub-Gaussian noise η_2 independent of D).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the reference-based early stopping framework, the three pseudo-reference algorithms, the theoretical support, and the empirical scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper states method-specific trade-offs in Section 3 and discusses theory and algorithm limitations in appendices.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions are stated with the theoretical results, and formal proofs or derivations are provided in appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 describes the datasets, noise models, baselines, metrics, and evaluation protocol; additional algorithmic details are in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the code after acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section specifies the tasks, datasets, noise models, optimizer settings where used, metrics, and oracle-gap definitions needed to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main tables report mean $\pm$ standard deviation across images, and the PSNR Gap metric reports the distance to the trajectory oracle.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments are not largely affected by the use of GPUs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is an algorithmic study using standard imaging data and does not involve prohibited data collection, human-subject intervention, or deployment decisions.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The current draft does not include a separate broader-impact discussion. The expected impacts are indirect and mainly concern improved reliability of image reconstruction methods.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models or datasets such as large generative models, language models, or unsafe scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: CC-BY 4.0

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset or model asset; the proposed algorithms are documented in the paper and appendices.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human-subject experiments, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core method, experiments, or originality of the research; any use for writing or editing does not require declaration under the question.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.