METROPOLIS-SSD: SYMMETRIC STOCHASTIC DENOISERS FOR CONVERGENT IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Denoiser-driven methods have become a powerful approach to iterative image reconstruction. Deep denoisers can achieve excellent reconstruction quality, but convergence of the resulting iterations is difficult to guarantee. Such guarantees typically require global properties such as nonexpansiveness, which are difficult to enforce for nonlinear denoisers. Linear denoisers offer a more tractable alternative, as their spectral properties can be analyzed and controlled directly. Their limited expressive power, however, can restrict reconstruction quality. A recent approach, NECTR, bridges this gap by separating nonlinear weight prediction from linear pixel aggregation, using a pretrained network to predict the weights. For quadratic data fidelity, NECTR guarantees convergence by enforcing the resulting linear denoiser ensuring that the resulting linear denoiser is stochasticity can easily be enforced by normalizing the predicted weights, imposing symmetry simultaneously is more challenging. We introduce Metro-SSD, which uses a simple Metropolis-type normalization to enforce both symmetry and stochasticity. Metro-SSD admits an exact matrix free implementation requiring only two aggregation passes, making it practical for large scale problems. Under a mild condition on the imaging forward model, we establish convergence guarantees for Metro-SSD with PnP-HQS and PnP-DRS. Using an attention-based network to predict the aggregation weights, Metro-SSD outperforms or matches the performance of state-of-the-art baselines in deblurring, superresolution and MRI.

1 INTRODUCTION

Many image reconstruction problems can be modeled as a

$$y = A\bar{x} + \epsilon,$$

where $y \in \mathbb{R}^m$ denotes the observed data, $\bar{x} \in \mathbb{R}^n$ the unknown image, $A \in \mathbb{R}^{m\times n}$ the imaging forward operator, and ϵ the measurement noise. This model covers deblurring, superresolution (SR), magnetic resonance imaging (MRI), and computed tomography (Elad et al., 2023; Bouman, 2022). Since A is often ill-conditioned or rank-deficient, classical model-based reconstruction formulates the inverse problem as

$$\min_{x\in\mathbb{R}^n} \; f(x) + g(x), \qquad f(x) := \frac{1}{2}\|Ax - y\|_2^2, \tag{1}$$

where f is the quadratic data-fidelity term and g is a regularizer that incorporates prior information about the image. Optimization problems of this form can be solved using splitting methods such as proximal gradient descent (PGD), half-quadratic splitting (HQS), Douglas-Rachford splitting (DRS), and the alternating direction method of multipliers (ADMM) (Geman & Yang, 1995; Eckstein & Bertsekas, 1992; Bauschke & Combettes, 2011). For example, the PGD iteration is

$$x_{k+1} = \mathrm{prox}_{\gamma g}\big(x_k - \gamma\nabla f(x_k)\big), \tag{2}$$

where prox_h denotes the proximal operator associated with a function h. The effectiveness of model-based depends strongly on the choice of g, and simple handcrafted regularizers often fail to capture the rich statistics of natural images (Elad et al., 2023).

Denoiser-driven regularization. The gradient step in (2) enforces data consistency but can introduce or amplify artifacts. The proximal step acts as a denoiser that suppresses these artifacts according to the prior encoded by g. Instead of going through a handcrafted regularizer and its proximal operator,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Plug-and-Play (PnP) methods directly replace the proximal step with a trained denoiser (Venkatakrishnan et al., 2013; Sreehari et al., 2016; Zhang et al., 2017). For example, the PnP variant of PGD (PnP-PGD) takes the form

$$x_{k+1} = \mathcal{D}\big(x_k - \gamma \nabla f(x_k)\big),$$

where $\mathcal{D} : \mathbb{R}^n \to \mathbb{R}^n$ denotes the denoising operator. Regularization-by-Denoising (RED) follows a different approach and uses the denoiser to define the regularization term implicitly, with the residual $x - \mathcal{D}(x)$ playing the role of its gradient (Romano et al., 2017; Reehorst & Schniter, 2018). A key advantage of PnP and RED is their modularity. The denoiser $\mathcal{D}$ can be trained independently of the forward operator and reused across different inverse problems without retraining (Zhang et al., 2021; Hurault et al., 2022a; Goujon et al., 2023; Pourya et al., 2025; Wei et al., 2025). In contrast, standard end-to-end models Guo et al. (2024); Liang et al. (2021) are typically trained for a particular forward model. More recent approaches seek to incorporate similar forms of modularity and forward-model adaptability of PnP methods into end-to-end reconstruction architectures (Terris et al., 2026; Potlapalli et al., 2023; Gossard & Weiss, 2024).

Convergent frameworks. Despite their empirical success, freely trained denoisers can cause PnP and RED to diverge (Reehorst & Schniter, 2018; Nair & Chaudhury, 2024). Existing convergent methods can be grouped into two categories. In one approach, the denoiser is associated with an explicit regularizer, whereby the resulting iterations can be interpreted as solving an optimization problem of the form (1). This can be achieved either by requiring the denoiser to be a proximal operator (Moreau, 1965; Gribonval & Nikolova, 2020; Yukawa & Yamada, 2025), or by requiring it to be a gradient step of a potential function (Cohen et al., 2021; Hurault et al., 2022a; Goujon et al., 2023; 2024; Shumaylov et al., 2024; Pourya et al., 2025; Hong et al., 2025). In either case, convergence of PnP or RED can be established using classical optimization theory (Bauschke & Combettes, 2011; Attouch & Bolte, 2009; Attouch et al., 2013). The flip side is that this restricts the class of neural denoisers to those that admit a proximal or potential-based interpretation.

In the second approach, PnP or RED is formulated as a fixed-point iteration of the form

$$x_{k+1} = \mathcal{T}(x_k), \tag{3}$$

where $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ is the update operator determined by the denoiser $\mathcal{D}$ and the splitting scheme. The explicit forms of $\mathcal{T}$ for different splitting schemes are given in Appendix A.

Convergence is then established by controlling properties of $\mathcal{T}$, often through global Lipschitz constraints on the denoiser (Ryu et al., 2019). Several works enforce such constraints directly at the network level using spectral normalization, orthogonal weight matrices, Lipschitz-constrained activations, or related Lipschitz-preserving constructions (Hertrich et al., 2021; Goujon et al., 2024; Ducotterd et al., 2024; Tanielian & Biau, 2021). These approaches provide global guarantees, while restricting the class of admissible network architectures and potentially limiting reconstruction performance. Other methods (Hurault et al., 2022b; Pesquet et al., 2021; Wei et al., 2024; 2025) adopt a less restrictive strategy by encouraging the desired Lipschitz behavior through training penalties evaluated on finite batches. These methods are less restrictive than globally constrained architectures but do not provide the same global guarantees.

Linear denoisers. A more tractable route to convergence guarantees is to use linear denoisers, whose spectral properties can be characterized and controlled directly (Sinha et al., 2025). In particular, under a mild condition on the forward operator, symmetry, stochasticity, and primitivity are sufficient to guarantee convergence. Purely linear denoisers, however, have limited expressive power. The recent NECTR framework (Sinha et al., 2026) combines linear aggregation with a neural network that predicts spatially adaptive weights. Within this framework, nonnegativity and stochasticity can be readily enforced through positive weights and normalization, but enforcing symmetry simultaneously is more difficult (Milanfar, 2013). NECTR therefore adopts the construction of Sreehari et al. (2016), which uses two-sided normalization, rescaling, and a diagonal correction. The construction also preserves positive semidefiniteness, which is needed for a proximal interpretation but not for fixed-point convergence. However, this additional requirement increases complexity and excludes symmetric stochastic denoisers with negative eigenvalues that satisfy the convergence conditions considered here.

Contributions. In this work, we develop a simpler way to enforce symmetry and stochasticity. Our contributions are:

- We introduce a new mechanism for constructing trainable symmetric stochastic denoisers using the classical Metropolis normalization (Xiao et al., 2005; Metropolis et al., 1953). The resulting Metro-SSD denoiser admits a matrix-free implementation, requiring only two aggregation passes.
- We characterize the spectral properties of Metro-SSD and establish convergence guarantees for splitting algorithms such as PGD, HQS and DRS.
- Using an attention-based weight-prediction network, we show that Metro-SSD achieves strong reconstruction performance, outperforming established baselines for image deblurring and MRI while remaining competitive for SR.

2 Trainable Linear Denoisers

The present work is motivated by the NECTR framework of Sinha et al. (2026), which combines nonlinear weight prediction with linear image aggregation. Specifically, a linear denoiser maps an image $x \in \mathbb{R}^n$ to $y = Wx$ through an aggregation matrix $W \in \mathbb{R}^{n\times n}$, with

$$y_i = \sum_{j=1}^{n} W_{ij} x_j. \tag{4}$$

When the weights are nonnegative and satisfy $We = e$, where $e \in \mathbb{R}^n$ is the all-ones vector, each output pixel is a weighted average of the input pixels. Assigning larger weights to similar pixels and smaller weights to dissimilar ones suppresses noise while preserving edges and other image structures. This principle underlies classical similarity-based filters (Buades et al., 2005; Tomasi & Manduchi, 1998; Dabov et al., 2007a).

NECTR framework. Classical similarity-based filters compute the weights W from the observed image using a prescribed rule. NECTR instead learns these weights from natural images by training a neural network to predict spatial weight maps from a reference image z, which serves as a surrogate for the unknown ground truth. Since predicting W entrywise would require generating and storing many pairwise pixel weights, NECTR represents W using image transformations and spatial weight maps, avoiding the explicit $n \times n$ matrix and enabling efficient prediction and application.

To make this representation precise, let $x : \Omega \to \mathbb{R}$ be an image on the periodic grid $\Omega = \mathbb{Z}_H \times \mathbb{Z}_L$, with $n = HL$ pixels. For $\tau \in \Omega$, define the translation operator

$$(T_\tau x)(i) = x(i - \tau) \qquad (i \in \Omega),$$

where the subtraction is modulo H and L. Let $\mathcal{S} \subset \Omega$ denote the set of translations used in the aggregation. For each $\tau \in \mathcal{S}$, define the spatial weight map $\omega_\tau : \Omega \to \mathbb{R}$ by

$$\omega_\tau(i) = W_{i,i-\tau} \qquad (i \in \Omega).$$

Then (4) can be expressed as

$$Wx = \sum_{\tau\in\mathcal{S}} D_\tau T_\tau x, \qquad D_\tau := \mathrm{Diag}(\omega_\tau). \tag{5}$$

Thus, translated copies of the input are weighted pointwise and then aggregated, without explicitly forming the matrix W.

NECTR generalizes the translation-based representation in (5) by replacing the translations with a finite set $\mathcal{G}$ of permutations of Ω. For each $\pi \in \mathcal{G}$, let $P_\pi : \mathbb{R}^n \to \mathbb{R}^n$ denote the corresponding permutation operator,

$$(P_\pi x)(i) = x(\pi^{-1}(i)) \qquad (i \in \Omega). \tag{6}$$

The aggregation operator then takes the form

$$W^z = \sum_{\pi\in\mathcal{G}} D_\pi P_\pi, \qquad \omega_\pi := \mathcal{N}_\theta(z, P_\pi z), \qquad D_\pi := \mathrm{Diag}(\omega_\pi), \tag{7}$$

where $\mathcal{N}_\theta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n_{++}$ is a positive-valued neural network that predicts the spatial weight map ω_π from a reference image z and its permuted copy. Thus, each permutation determines which pixels are aggregated, while the predicted weight map determines their contributions. For brevity, once the reference image z is fixed, we write W in place of W^z.

Symmetry and stochasticity. As the operator W in (7) is generally not symmetric, NECTR enforces symmetry by pairing each term with its transpose and defining

$$K = \sum_{\pi \in \mathcal{G}} \left(D_\pi P_\pi + P_{\pi^{-1}} D_\pi \right). \tag{8}$$

Since $(D_\pi P_\pi)^\top = P_{\pi^{-1}} D_\pi$, the operator K is symmetric, but not necessarily stochastic. A simple way to enforce stochasticity is through row normalization. Specifically, letting $d = Ke$ and $\Delta = \mathrm{Diag}(d)$, the operator $W := \Delta^{-1} K$ satisfies $We = e$. However, the left scaling by Δ^{-1} generally destroys symmetry. The challenge is therefore to enforce stochasticity while preserving symmetry using only simple corrections.

To address this, NECTR adopts the three-stage construction of Sreehari et al. (2016). It first applies the symmetric normalization

$$S = \Delta^{-1/2} K \Delta^{-1/2},$$

followed by the diagonal correction

$$H = \alpha S + \mathrm{Diag}(e - \alpha^{-1}\, Se), \qquad \alpha := \max_{1 \leqslant i \leqslant n} (Se)_i.$$

The resulting operator H is symmetric, nonnegative, and stochastic. To compute Hx, we need three aggregation passes to obtain Ke, Se, and Sx.

The motivation behind the above construction of Sreehari et al. (2016) is to preserve positive semidefiniteness: if K is positive semidefinite, then so is H. This gives H a proximal interpretation (Moreau, 1965) and allows convergence to be established through standard optimization theory. For fixed-point convergence, however, H need not be positive semidefinite. Dropping this requirement broadens the class of admissible denoisers and simplifies the normalization mechanism.

3 METROPOLIS-SSD

We propose **Metro-SSD**, which adapts the classical Metropolis normalization to the trainable linear aggregation framework developed above. This allows symmetry and stochasticity to be enforced directly, without requiring positive semidefiniteness or the three-stage normalization used in NECTR. The construction builds on the convergence result of Sinha et al. (2025), which shows that symmetry, stochasticity, and primitivity are sufficient to make the associated PnP operators contractive under a mild condition on the forward model. Contractivity then guarantees linear convergence of the iterates to a unique fixed point.

3.1 METROPOLIS NORMALIZATION

We show how symmetry and stochasticity can be enforced simultaneously using the Metropolis scheme (Xiao et al., 2005). Starting from the symmetric aggregation operator K in (8), define its degree vector by $d = Ke$. The Metropolis rule rescales each entry K_{ij} by the larger of the two associated degrees,

$$(K_{\mathrm{M}})_{ij} = \frac{K_{ij}}{\max(d_i, d_j)} \qquad (1 \leqslant i, j \leqslant n). \tag{9}$$

Since the denominator is symmetric in i and j, this scaling preserves symmetry and nonnegativity. Moreover,

$$(K_{\mathrm{M}} e)_i = \sum_{j=1}^{n} \frac{K_{ij}}{\max(d_i, d_j)} \leqslant \frac{1}{d_i} \sum_{j=1}^{n} K_{ij} = 1.$$

The remaining row mass is assigned to the diagonal by defining

$$W_{\mathrm{M}} = K_{\mathrm{M}} + \mathrm{Diag}(e - K_{\mathrm{M}} e).$$

Since $K_{\mathrm{M}} e \leqslant e$, the diagonal correction is nonnegative and preserves symmetry, while $W_{\mathrm{M}} e = e$. Hence W_{M} is symmetric and stochastic.

The entrywise rule in (9) appears to require explicit access to the entries of K. This can be avoided by exploiting the permutation representation in (8). Indeed, for $\pi \in \mathcal{G}$, the term $D_\pi P_\pi$ couples pixel

i with $\pi^{-1}(i)$, so the corresponding Metropolis denominator can be written as

$$\max\left(d_i, d_{\pi^{-1}(i)}\right) = \max\left(d_i, (P_\pi d)_i\right) = \big(\max(d, P_\pi d)\big)_i,$$

where the maximum is taken componentwise. Accordingly, define

$$M_\pi = \mathrm{Diag}\left(\max(d, P_\pi d)\right).$$

Then

$$K_{\mathrm{M}} = \sum_{\pi \in \mathcal{G}} \left(M_\pi^{-1} D_\pi P_\pi + P_{\pi^{-1}} D_\pi M_\pi^{-1}\right). \tag{10}$$

This exactly implements the scaling in (9) without forming or storing K. The final denoiser is

$$\text{(Metro-SSD)} \qquad W_{\mathrm{M}} = K_{\mathrm{M}} + \mathrm{Diag}(e - K_{\mathrm{M}} e). \tag{11}$$

The remaining requirement for convergence is primitivity. We say that W_{M} is primitive if $(W_{\mathrm{M}})^p > 0$ entrywise for some $p \geqslant 1$. For Metro-SSD, primitivity follows when the permutation set $\mathcal{G}$ is transitive and contains the identity permutation. By transitivity of $\mathcal{G}$, we mean that for any $i, j \in \Omega$, there exist $m \geqslant 1$ and $\pi_1, \ldots, \pi_m \in \mathcal{G} \cup \mathcal{G}^{-1}$ such that

$$(\pi_m \circ \cdots \circ \pi_1)(i) = j.$$

The resulting properties are summarized below, with the proof deferred to Appendix B.

Proposition 1 (Properties of Metro-SSD). *Let $\mathcal{G}$ be a nonempty set of permutations of Ω, and suppose that $\omega_\pi > 0$ for all $\pi \in \mathcal{G}$. Then the Metro-SSD operator W_{M} in (11) is well-defined, symmetric, nonnegative, and stochastic. In addition, if $\mathcal{G}$ is transitive and contains the identity permutation, then W_{M} is primitive.*

The conditions in Proposition 1 can be readily enforced by design. For example, including the identity together with the one-pixel up, down, left, and right translations makes $\mathcal{G}$ transitive.

3.2 Neural Weight Predictor

We next specify the neural network used to predict the aggregation weights from the reference image. Similar to NECTR (Sinha et al., 2026), we use a fixed set $\mathcal{G}$ of cyclic translations, which define a structured local neighborhood for denoising and allow efficient aggregation through spatial shifts. We use integer offsets from $-R$ to R along each image axis, with $R = 5$, giving $(2R+1)^2 = 121$ shifts. Rather than predicting a separate weight map for every shift, we pair each shift with its inverse and predict one map per pair, with the inverse contribution supplied by the symmetric construction (Sinha et al., 2026, Algorithm 1). This reduces the number of predicted maps to $J = 2R^2 + 2R + 1 = 61$.

The network takes a reference image $z \in \mathbb{R}^n$ as input and jointly predicts the J strictly positive spatial weight maps, one for the identity shift and one for each inverse pair,

$$\mathcal{N}_\theta : \mathbb{R}^n \longrightarrow (\mathbb{R}^n_{++})^J, \qquad \mathcal{N}_\theta(z) = \big(\omega_1(z), \ldots, \omega_J(z)\big).$$

We construct K in (8) using these inverse pairs, as described in Appendix C, and apply the Metropolis normalization (10)–(11) to obtain W^z_{M}.

For the network architecture, we use a DRUNet-based encoder-decoder (Zhang et al., 2021) to extract full-resolution features from the reference image z together with a spatially constant noise-level map σe. The features and their query–key projections are computed once. For each shift, multi-head attention (Vaswani et al., 2017; Wang et al., 2018) scores a slice of the padded key tensor against the central queries. A softmax over the resulting shift scores produces strictly positive affinities. The features and query-key projections are normalized using LayerNorm (Ba et al., 2016; Henry et al., 2020), and a head-mixing layer with softplus-activated coefficients preserves positivity of the final weight maps. The DRUNet backbone uses channel widths $(64, 128, 256, 384)$ across four scales, with two residual blocks per scale, ReLU activations, and skip connections, producing a 192-channel feature representation. An eight-head attention module with query/key dimension 96 then predicts the weight maps. For multichannel images, we predict channel-specific maps, as illustrated in Figure 4. Further implementation details are provided in Appendix D.1.

We use a standard denoiser training protocol following Zhang et al. (2021). We train the resulting Metro-SSD operator W_{M} as a single-step denoiser by minimizing the empirical mean squared error (MSE) between its outputs on noisy image patches and the corresponding denoised targets. Training details are provided in Appendix D.2.

4 PLUG-AND-PLAY RECONSTRUCTION

We now use the trained Metro-SSD denoiser within PnP algorithms for image reconstruction. We first describe the algorithms considered in this work and the warm-up strategy used to fix the reference image. We then establish linear convergence guarantees for the resulting iterations. Finally, we present reconstruction results for image deblurring, superresolution, and MRI, together with comparisons against state-of-the-art methods.

4.1 ALGORITHMS

In this paper, we consider PnP-HQS and PnP-DRS (Zhang et al., 2021; Hurault et al., 2024), as they provide the best reconstruction performance in our experiments. However, we note that Metro-SSD can also be incorporated into other denoiser-driven reconstruction schemes, including PnP-PGD (Ryu et al., 2019; Hurault et al., 2022b) and RED (Romano et al., 2017; Reehorst & Schniter, 2018), for which analogous convergence guarantees can be established.

For a fixed Metro-SSD operator W_{M}, both PnP-HQS and PnP-DRS generate iterates of the form (3), with the fixed-point operator $\mathcal{T}$ determined by the splitting scheme. For PnP-HQS with $\rho > 0$,

$$\mathcal{T}_\rho = W_{\mathrm{M}} \circ \mathrm{prox}_{\rho f}\,. \tag{12}$$

For PnP-DRS, we use the half-averaged denoiser $(I + W_{\mathrm{M}})/2$, whose reflected operator is W_{M}. Thus, for $\mu > 0$ and $\beta \in (0, 1]$,

$$\mathcal{T}_{\mu,\beta} = (1-\beta)I + \beta(2\,\mathrm{prox}_{\mu f} - I) \circ W_{\mathrm{M}}. \tag{13}$$

Following Sinha et al. (2026), we update the reference image used to construct W_{M} during an initial warm-up phase. Let $Q \geqslant 1$ denote the number of warm-up iterations. For PnP-HQS, starting from $x_0 \in \mathbb{R}^n$, we use the current reconstruction as the reference and perform

$$x_{k+1} = \big(W_{\mathrm{M}}^{x_k} \circ \mathrm{prox}_{\rho f}\big)(x_k) \qquad (0 \leqslant k \leqslant Q-1). \tag{14}$$

Thus, the predicted weight maps are updated as the reconstruction evolves. After Q iterations, we set the reference to x_Q and freeze the denoiser at $W_{\mathrm{M}} = W_{\mathrm{M}}^{x_Q}$. The subsequent iterations are

$$x_{k+1} = \big(W_{\mathrm{M}} \circ \mathrm{prox}_{\rho f}\big)(x_k) \qquad (k \geqslant Q).$$

The same warm-up principle applies to PnP-DRS and other PnP or RED algorithms. In our experiments, we use a modified procedure that collects the inputs to the denoiser as candidate reference images, rather than the outer iterates, and selects a reference from these candidates. We describe the implementation in Appendix D.3 and examine the warm-up length and reference-selection scheme in Appendix E.2. Once the warm-up phase ends, W_{M} is fixed, and we can analyze the resulting PnP iterations using the linear-denoiser theory developed in Sinha et al. (2025).

4.2 CONVERGENCE GUARANTEES

By construction, the frozen Metro-SSD operator W_{M} is symmetric, stochastic, and primitive. Consequently, 1 is a simple eigenvalue with eigenvector e, while all remaining eigenvalues lie in $(-1, 1)$. Hence, $\mathrm{span}\{e\}$ is the only subspace on which W_{M} has unit gain. For quadratic data fidelity, $\mathrm{prox}_{\rho f}$ has unit gain precisely on $\mathrm{null}(A)$. The condition $Ae \neq 0$, equivalently $\mathrm{span}\{e\} \cap \mathrm{null}(A) = \{0\}$, therefore rules out a common unit-gain direction. This yields strict contractivity of the composite PnP operator (Sinha et al., 2025). The precise result is stated below, with the proof given in Appendix B.

Theorem 1 (Contractive PnP). *Let $W_{\mathrm{M}} = W_{\mathrm{M}}^z$ be the Metro-SSD operator associated with a fixed reference image z, and let $\mathcal{T}_\rho$ and $\mathcal{T}_{\mu,\beta}$ be the operators defined in* (12) *and* (13)*. If $Ae \neq 0$, then $\mathcal{T}_\rho$ is contractive for every $\rho > 0$, and $\mathcal{T}_{\mu,\beta}$ is contractive for every $\mu > 0$ and $\beta \in (0, 1]$. Consequently, the corresponding fixed-point iterations converge linearly to unique fixed points.*

The analysis extends naturally to PnP-PGD and RED. For PnP-PGD (Ryu et al., 2019; Hurault et al., 2022b), the operator

$$\mathcal{T}_\gamma = W_{\mathrm{M}} \circ (I - \gamma \nabla f)$$

is contractive for $0 < \gamma < 2/\|A\|_2^2$. For Metro-SSD-based RED (Romano et al., 2017), the corresponding iterations can be viewed as gradient descent applied to

$$\min_{x \in \mathbb{R}^n} \; f(x) + \lambda g(x), \qquad g(x) := \frac{1}{2}\, x^\top (I - W_{\mathrm{M}})x.$$

Table 1: PSNR (dB) comparison on the Knee-MRI dataset at noise level $\nu = 0.002$. Best results are **bold**, second best underlined, and third best *italicized*.

Method	2-fold single-coil		4-fold single-coil		4-fold multi-coil		8-fold multi-coil		Avg
	PD	PDFS	PD	PDFS	PD	PDFS	PD	PDFS	
Zero-fill ($A^\top y$)	35.87	36.59	29.95	32.01	30.26	32.26	26.26	29.52	31.59
ADMM+TV (Rudin et al., 1992)	41.69	39.28	35.14	34.49	40.70	39.02	34.83	34.85	37.50
ADMM+MMO (Pesquet et al., 2021)	44.08	40.61	38.37	36.17	42.76	40.08	38.12	36.66	39.61
HQS+LipDSNN (Ducotterd et al., 2024)	38.54	36.57	34.63	33.97	37.82	36.00	34.59	34.19	35.79
GSPnP-HQS (Hurault et al., 2022a)	43.94	40.58	38.25	*36.63*	42.99	*40.76*	*38.28*	37.35	39.85
ProxPnP-DRS (Hurault et al., 2022b; 2024)	*44.15*	40.80	*38.47*	36.68	*43.04*	40.85	37.93	36.81	*39.84*
AGD+CRR (Goujon et al., 2023)	43.62	40.36	37.19	35.73	41.88	40.05	37.68	36.85	39.17
SAGD+WCRR (Goujon et al., 2024)	43.93	40.61	38.36	36.33	43.06	40.32	38.25	36.71	39.69
CG+DEAL (Pourya et al., 2025)	44.22	*40.66*	**39.19**	36.55	43.02	40.26	**38.80**	*36.99*	39.85
DRS+Metro-SSD (Ours)	**44.37**	**40.86**	38.61	**36.74**	**43.25**	**40.87**	38.34	**37.47**	**40.07**

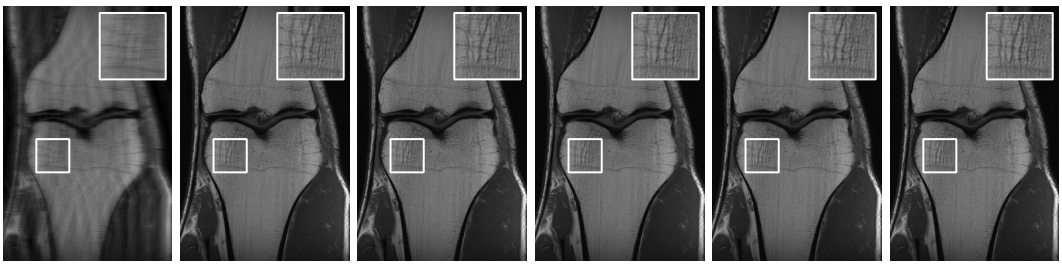
Since W_{M} is symmetric with eigenvalues in $(-1, 1]$, the function g is convex, while primitivity gives $\mathrm{null}(I - W_{\mathrm{M}}) = \mathrm{span}\{e\}$. Hence, if $Ae \neq 0$, then $f + \lambda g$ is strongly convex for every $\lambda > 0$, yielding convergence guarantees for the RED iterations (Sinha & Chaudhury, 2026). The same structure also provides linear convergence guarantees for proximal RED algorithms (Reehorst & Schniter, 2018).

The non-cancellation condition $Ae \neq 0$ holds for many common inverse problems, including inpainting, image deblurring, superresolution, computed tomography, and MRI. In inpainting, $Ae \neq 0$ whenever at least one pixel is observed. Normalized blur kernels preserve constants, so the condition holds for deblurring and remains valid after downsampling in superresolution. In tomography, a nonzero constant image produces nonzero line integrals under any nontrivial acquisition geometry. For MRI, if the sampling mask contains the DC frequency, then $Ae \neq 0$ whenever at least one coil sensitivity map has nonzero spatial mean. This holds automatically in the single-coil case with unit sensitivity. We also verified $Ae \neq 0$ numerically for the multi-coil MRI operators used in our experiments (Goujon et al., 2023). Thus, the convergence guarantees in Theorem 1 apply to these imaging settings.

4.3 Reconstruction Results

We evaluate the empirical performance of Metro-SSD across different reconstruction tasks and degradation settings. Our experiments cover image deblurring, superresolution, and MRI, with comparisons against established PnP baselines.

For deblurring and superresolution, we use PnP-HQS on the CBSD10 dataset used in Hurault et al. (2022a). Deblurring is evaluated using nine standard blur kernels at noise levels $\nu \in \{0.03, 0.05\}$, while superresolution is evaluated with four isotropic Gaussian blur kernels and scale factors $s \in \{2, 3\}$ at the same noise levels. For MRI, we use PnP-DRS on proton-density (PD) and fat-suppressed proton-density (PDFS) knee images, considering both single- and multi-coil settings from the FastMRI-based dataset of Goujon et al. (2023). We set $\nu = 0.002$ and use Cartesian Fourier subsampling with acceleration $M_{\mathrm{acc}} \in \{2, 4, 8\}$ and center fraction $M_{\mathrm{cf}} = 0.32/M_{\mathrm{acc}}$. We run 100 PnP iterations for deblurring and superresolution and 500 for MRI, including the warm-up iterations.



zero-fill ($A^\top y$) 29.19 dB | ProxPnP-DRS 43.59 dB | GSPnP-HQS 43.48 dB | DEAL 43.50 dB | Metro-SSD 43.67 dB | clean

Figure 1: Knee-MRI reconstruction for the 4-fold multicoil PD setting with noise level 0.002.

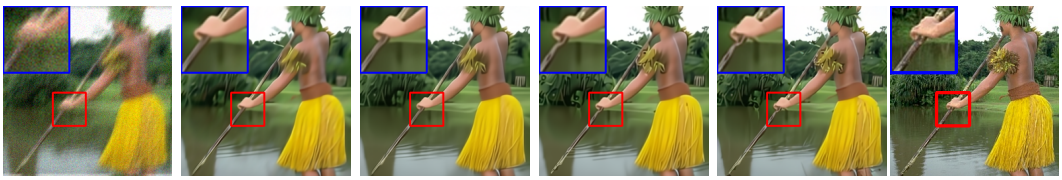
Table 2: PSNR (dB) comparison of deblurring methods on CBSD10 across blur kernels and noise levels $\nu \in \{0.03, 0.05\}$. Best scores are **bold**, second best underlined, and third best *italicized*.

ν	Method										Avg
0.03	Observed	21.48	20.96	21.68	17.60	22.06	17.92	18.89	18.76	23.48	20.32
	DPIR+DRUNet (Zhang et al., 2021)	29.22	**28.87**	29.01	**28.60**	30.06	*30.06*	**29.28**	28.83	27.84	**29.09**
	GSPnP-HQS (Hurault et al., 2022a)	*29.17*	28.85	**29.14**	28.58	29.94	**30.10**	**29.28**	**28.90**	**27.85**	**29.09**
	ProxPnP-DRS (Hurault et al., 2022b; 2024)	28.57	28.59	28.91	28.25	30.11	29.87	29.07	28.67	*27.79*	28.87
	ProxPnP-PGD (Hurault et al., 2022b; 2024)	28.97	28.62	28.79	28.35	30.17	30.03	29.02	28.67	*27.79*	28.94
	ProxPnP-αPGD (Hurault et al., 2023; 2024)	28.75	28.72	*29.03*	*28.42*	*30.13*	29.95	*29.13*	28.79	**27.85**	28.97
	Ann-SNORE (Renaud et al., 2024)	28.58	28.30	28.58	27.93	29.46	29.38	28.54	28.26	27.35	28.49
	IHQS+SPC-DRUNet (Wei et al., 2024)	27.85	27.28	27.61	26.88	29.12	29.04	27.90	27.31	26.95	27.77
	ADMM+CoCo-DRUNet (Wei et al., 2025)	28.74	28.29	28.46	27.98	29.81	29.70	28.71	28.28	27.55	28.61
	CG+DEAL (Pourya et al., 2025)	29.11	*28.81*	*29.03*	28.58	29.95	29.79	28.98	28.71	**27.85**	*28.98*
	HQS+NECTR (Sinha et al., 2026)	28.23	27.89	28.32	27.57	29.48	29.25	28.46	28.01	27.44	28.29
	HQS+Metro-SSD (Ours)	**29.23**	28.85	29.06	28.58	**30.18**	30.07	29.15	*28.72*	27.78	29.07
0.05	Observed	20.45	20.03	20.60	17.14	20.90	17.43	18.28	18.17	22.00	19.44
	DPIR+DRUNet (Zhang et al., 2021)	*27.39*	*27.15*	*27.59*	26.90	28.50	*28.28*	**27.73**	**27.27**	**27.15**	27.55
	GSPnP-HQS (Hurault et al., 2022a)	27.36	27.16	27.61	**26.91**	28.55	28.27	27.32	26.99	26.71	27.43
	ProxPnP-DRS (Hurault et al., 2022b; 2024)	27.08	27.10	27.46	26.71	28.52	28.20	27.58	27.21	27.05	27.43
	ProxPnP-PGD (Hurault et al., 2022b; 2024)	27.29	27.09	27.39	26.73	28.58	28.34	*27.61*	*27.23*	*27.04*	*27.48*
	ProxPnP-αPGD (Hurault et al., 2023; 2024)	27.03	27.16	27.50	26.77	28.37	28.12	27.48	**27.27**	26.99	27.41
	Ann-SNORE (Renaud et al., 2024)	26.75	26.62	27.04	26.37	27.40	27.31	26.79	26.71	26.62	26.84
	IHQS+SPC-DRUNet (Wei et al., 2024)	26.73	26.20	26.67	25.91	27.90	27.76	26.92	26.44	26.54	26.79
	ADMM+CoCo-DRUNet (Wei et al., 2025)	27.12	26.95	27.28	26.61	28.27	28.05	27.39	26.99	26.91	27.29
	CG+DEAL (Pourya et al., 2025)	27.47	27.06	27.43	26.83	28.41	28.20	27.51	27.18	27.05	27.46
	HQS+NECTR (Sinha et al., 2026)	26.42	26.28	26.87	25.95	27.82	27.47	26.88	26.49	26.62	26.75
	HQS+Metro-SSD (Ours)	**27.52**	**27.29**	**27.62**	**26.91**	**28.66**	**28.42**	27.63	27.24	27.05	**27.60**

For a fair comparison, we adopt standard protocols and benchmark Metro-SSD against state-of-the-art Gaussian denoiser-based PnP methods under similar reconstruction settings. For this reason, we do not include diffusion and flow based methods. This broad range of baselines include empirical methods such as DPIR (Zhang et al., 2021) and several methods with convergence guarantees, including MMO (Pesquet et al., 2021), GSPnP (Hurault et al., 2022a), ProxPnP (Hurault et al., 2022b; 2023; 2024), Ann-SNORE (Renaud et al., 2024), (W)CRR (Goujon et al., 2023; 2024), LipDSNN (Ducotterd et al., 2024), DEAL (Pourya et al., 2025), SPC- and CoCo-DRUNets (Wei et al., 2024; 2025), and NECTR (Sinha et al., 2026). Among these methods, MMO, ProxDRUNet, and SPC- and CoCo-DRUNets enforce the relevant constraints through training-time Jacobian regularization and DEAL enforces partial guarantees. Consequently, these constraints are not guaranteed to hold globally and may be violated by test-time iterates outside the training distribution (Goujon et al., 2024; Sinha et al., 2026).

For all methods, we initialize the PnP iterations with the observed image for deblurring, bicubic interpolation of the low-resolution observation for superresolution, and the zero-filled reconstruction $A^\top y$ for MRI. For Metro-SSD, we also set the initial reference to $z = x_0$. A warm-up of $Q = 150$ is used for MRI, while a warm-up of $Q = 30$ iterations is used for deblurring and superresolution. Numerical results in Appendix F show that the PSNR stabilizes, successive-iterate differences decay, and different initializations converge to the same fixed point. Additional experiments and ablations in Appendix E examine standalone denoising performance, further reconstruction examples, and the influence of the warm-up duration and implementation choices.

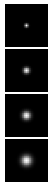
Summary. The results in Tables 1 to 3 show a consistent pattern across the three reconstruction tasks. Metro-SSD achieves the highest average PSNR in deblurring and MRI, while remaining competitive

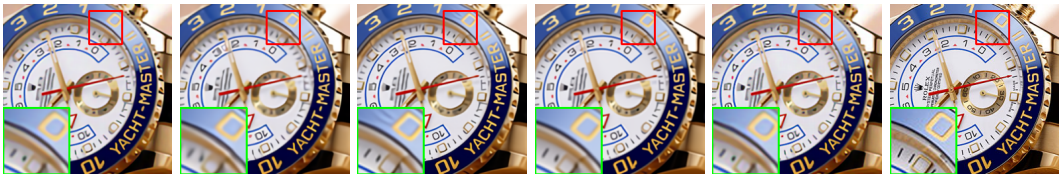


blurry 20.52 dB | ProxPnP-PGD 25.67 dB | GSPnP-HQS 25.95 dB | DPIR 25.77 dB | Metro-SSD 26.11 dB | clean

Figure 2: Motion deblurring results on the "man" image using kernel 2 at noise level 0.05.

Table 3: PSNR (dB) comparison of superresolution methods on CBSD10 for scale factors $s \in \{2, 3\}$ and noise levels $\nu \in \{0.03, 0.05\}$. Best are **bold**, second best underlined, and third best *italicized*.

Kernels	Method	$s = 2$		$s = 3$		Avg
		$\nu = 0.03$	$\nu = 0.05$	$\nu = 0.03$	$\nu = 0.05$	
	Bicubic	23.74	22.49	22.28	21.32	22.46
	DPIR+DRUNet (Zhang et al., 2021)	27.35	**26.49**	*25.87*	25.13	**26.21**
	GSPnP-HQS (Hurault et al., 2022a)	27.09	26.12	25.82	24.66	25.92
	ProxPnP-DRS (Hurault et al., 2022b; 2024)	27.31	*26.44*	*25.87*	25.16	26.20
	ProxPnP-PGD (Hurault et al., 2022b; 2024)	27.32	26.46	*25.87*	*25.14*	26.20
	ProxPnP-αPGD (Hurault et al., 2023; 2024)	*27.34*	26.34	**25.95**	**25.19**	**26.21**
	Ann-SNORE (Renaud et al., 2024)	26.87	25.86	25.52	24.92	25.79
	IHQS+SPC-DRUNet (Wei et al., 2024)	26.61	25.98	25.22	24.73	25.64
	ADMM+CoCo-DRUNet (Wei et al., 2025)	26.94	26.19	25.37	24.87	25.84
	CG+DEAL (Pourya et al., 2025)	27.22	26.33	25.88	25.05	26.12
	HQS+NECTR (Sinha et al., 2026)	26.77	25.86	25.21	24.49	25.58
	HQS+Metro-SSD (Ours)	**27.36**	26.46	25.83	24.98	*26.16*



bicubic (16.25 dB) | ProxPnP-αPGD (20.28 dB) | GSPnP-HQS (21.18 dB) | DPIR (20.25 dB) | Metro-SSD (20.43 dB) | clean

Figure 3: Results of 2× superresolution of "watch" with kernel 3 and noise level $\nu = 0.03$. The zoomed insets show some visual artifacts for GSPnP despite its strong PSNR.

with the strongest baselines in superresolution. It also consistently improves over NECTR, indicating that the simpler Metropolis normalization does not sacrifice reconstruction quality. The visual comparisons in Figures 1 to 3 are consistent with the quantitative results.

5 Conclusion

We introduced Metro-SSD, a trainable linear denoiser obtained by applying Metropolis normalization to learned aggregation weights. The neural predictor uses a reference image to generate spatially adaptive weights, which are converted into a symmetric and stochastic operator that is also primitive under strictly positive weights and an appropriate permutation set. After the warm-up phase, the reference is fixed and the resulting denoiser remains linear throughout the subsequent reconstruction. Thus, image adaptivity enters through the learned construction of the operator, while its fixed linear structure enables rigorous convergence analysis. These properties are enforced exactly without forming the underlying matrix, using only two aggregation passes, and do not require positive semidefiniteness. Under the mild non-cancellation condition that the forward operator does not annihilate the constant image, they yield contractive PnP operators and a strongly convex RED objective, with corresponding linear convergence guarantees.

Experiments on deblurring, superresolution, and MRI indicate that this structural restriction does not come at the expense of reconstruction quality. Metro-SSD consistently improves on the closely related NECTR construction and remains competitive with recent PnP methods, including methods based on more flexible nonlinear denoisers, across a range of degradation and noise levels.

The present theory is restricted to quadratic data fidelity and to the fixed-reference phase of the reconstruction. Extending the analysis to more general data fidelity terms and to iterations with an evolving learned operator are natural directions for future work.

AI USE STATEMENT

All primary research activities, including conceptualization, theoretical development, experimental design, data analysis, and the preparation of the original manuscript, were carried out by the human authors. Generative AI tools were used only as auxiliary aids for limited rephrasing to improve clarity and presentation, for generating preliminary versions of selected diagrams, and for assisting with portions of the implementation code. All AI-generated content was carefully reviewed, revised, and validated by the authors. The authors take full intellectual responsibility for the methodology, results, interpretations, and claims presented in this work.

REPRODUCIBILITY STATEMENT

We describe the method, experimental setup, training, datasets, and evaluation protocols in the main paper and appendix. Complete proofs are provided in the appendix. The source code is included as supplementary material and will be released publicly upon acceptance.

REFERENCES

Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, 2009.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical programming*, 137(1):91–129, 2013.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

Charles A Bouman. *Foundations of Computational Imaging: A Model-Based Approach*. SIAM, 2022.

Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. *Proc. CVPR*, 2:60–65, 2005.

Regev Cohen, Yochai Blau, Daniel Freedman, and Ehud Rivlin. It has potential: Gradient-driven denoisers for convergent solutions to inverse problems. *Proc. NeurIPS*, pp. 18152–18164, 2021.

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(8):2080–2095, 2007a.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. In *IEEE International Conference on Image Processing*, volume 1, pp. I – 313–I – 316, 2007b.

Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. *Proc. ECCV*, pp. 391–407, 2016.

Stanislas Ducotterd, Alexis Goujon, Pakshal Bohra, Dimitris Perdios, Sebastian Neumayer, and Michael Unser. Improving lipschitz-constrained neural networks by learning activation functions. *Journal of Machine Learning Research*, 25(65):1–30, 2024.

Jonathan Eckstein and Dimitri P. Bertsekas. On the douglas–rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55: 293–318, 1992.

Michael Elad, Bahjat Kawar, and Gregory Vaksman. Image denoising: The deep learning revolution and beyond—a survey paper. *SIAM J. Imaging Sci.*, 16(3):1594–1654, 2023.

Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Process.*, 4(7):932–946, 1995.

Alban Gossard and Pierre Weiss. Training adaptive reconstruction networks for blind inverse problems. *SIAM Journal on Imaging Sciences*, 17(2):1314–1346, 2024.

Alexis Goujon, Sebastian Neumayer, Pakshal Bohra, Stanislas Ducotterd, and Michael Unser. A neural-network-based convex regularizer for inverse problems. *IEEE Trans. Comput. Imaging*, 9: 781–795, 2023.

Alexis Goujon, Sebastian Neumayer, and Michael Unser. Learning weakly convex regularizers for convergent image-reconstruction algorithms. *SIAM J. Imaging Sci.*, 17(1):91–115, 2024.

Rémi Gribonval and Mila Nikolova. A characterization of proximity operators. *Journal of Mathematical Imaging and Vision*, 62(6):773–789, 2020. ISSN 0924-9907.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. *Proc. ECCV*, pp. 222–241, 2024.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. *Proc. EMNLP*, pp. 4246–4253, 2020.

Johannes Hertrich, Sebastian Neumayer, and Gabriele Steidl. Convolutional proximal neural networks and plug-and-play algorithms. *Linear Algebra Appl.*, 631:203–234, 2021.

Tao Hong, Zhaoyi Xu, Se Young Chun, Luis Hernandez-Garcia, and Jeffrey A Fessler. Convergent complex quasi-newton proximal methods for gradient-driven denoisers in compressed sensing mri reconstruction. *IEEE transactions on computational imaging*, 2025.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. *Proc. ICLR*, 2022a.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization. *Proc. ICML*, pp. 9483–9505, 2022b.

Samuel Hurault, Antonin Chambolle, Arthur Leclaire, and Nicolas Papadakis. A Relaxed Proximal Gradient Descent Algorithm for Convergent Plug-and-Play with Proximal Denoiser. *Proc. SSVMCV*, 2023.

Samuel Hurault, Antonin Chambolle, Arthur Leclaire, and Nicolas Papadakis. Convergent Plug-and-Play with Proximal Denoiser and Unconstrained Regularization Parameter. *Journal of Mathematical Imaging and Vision*, 66:616–638, 2024. ISSN 0924-9907, 1573-7683.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. *Proc. ICCV Workshops*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *Proc. CVPR Workshops*, pp. 1132–1140, 2017.

Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2017.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proc. ICCV*, 2:416–423, 2001.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

P. Milanfar. A tour of modern image filtering: new insights and methods, both practical and theoretical. *IEEE Signal Process. Mag.*, 30(1):106–128, 2013.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

Pravin Nair and Kunal N Chaudhury. Averaged deep denoisers for image regularization. *J. Math. Imaging Vis.*, 66(3):362–379, 2024.

Jean-Christophe Pesquet, Audrey Repetti, Matthieu Terris, and Yves Wiaux. Learning maximally monotone operators for image recovery. *SIAM J. Imaging Sci.*, 14(3):1206–1237, 2021.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in neural information processing systems*, 36:71275–71293, 2023.

Mehrsa Pourya, Erich Kobler, Michael Unser, and Sebastian Neumayer. DEALing with image reconstruction: Deep attentive least squares. *Proc. ICML*, 2025.

Edward T Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE Trans. Comput. Imaging*, 5(1):52–67, 2018.

Marien Renaud, Jean Prost, Arthur Leclaire, and Nicolas Papadakis. Plug-and-Play image restoration with Stochastic deNOising REgularization. *Proc. ICML*, 2024.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by Denoising (RED). *SIAM J. Imaging Sci.*, 10(4):1804–1844, 2017.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D: Nonlinear Phenom.*, 60(1-4):259–268, 1992.

Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. *Proc. ICML*, pp. 5546–5557, 2019.

Zakhar Shumaylov, Jeremy Budd, Subhadip Mukherjee, and Carola-Bibiane Schönlieb. Weakly Convex Regularisers for Inverse Problems: Convergence of Critical Points and Primal-Dual Optimisation. *Proc. ICML*, pp. 45286–45314, 2024. ISSN 2640-3498.

Arghya Sinha and Kunal N. Chaudhury. Strong convexity of (kernel) Laplacian regularization. *Proc. ICASSP*, pp. 11197–11201, 2026.

Arghya Sinha, Bhartendu Kumar, Chirayu D. Athalye, and Kunal N. Chaudhury. Linear convergence of plug-and-play algorithms with kernel denoisers. *IEEE Trans. Signal Process.*, 73:2646–2659, 2025.

Arghya Sinha, Aditya Banerjee, Trishit Mukherjee, and Kunal Narayan Chaudhury. Trainable nonexpansive denoisers for contractive image reconstruction. *Proc. ICML*, 2026.

Suhas Sreehari, S Venkat Venkatakrishnan, Brendt Wohlberg, Gregery T Buzzard, Lawrence F Drummy, Jeffrey P Simmons, and Charles A Bouman. Plug-and-play priors for bright field electron tomography and sparse interpolation. *IEEE Trans. Comput. Imaging*, 2(4):408–423, 2016.

Ugo Tanielian and Gerard Biau. Approximating lipschitz continuous functions with groupsort neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 442–450. PMLR, 2021.

Matthieu Terris, Samuel Hurault, Maxime Song, and Julián Tachella. Reconstruct anything model a lightweight general model for computational imaging. *Proc. ICLR*, 2026.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, et al. NTIRE 2017 challenge on single image super-resolution: Methods and results. *Proc. CVPR Workshops)*, pp. 1110–1121, 2017.

Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. *Proc. ICCV*, pp. 839–846, 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proc. NeurIPS*, 2017.

Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-Play priors for model based reconstruction. *Proc. IEEE GlobalSIP*, pp. 945–948, 2013.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *Proc. CVPR*, pp. 7794–7803, 2018.

Deliang Wei, Peng Chen, and Fang Li. Learning pseudo-contractive denoisers for inverse problems. *Proc. ICML*, 2024.

Deliang Wei, Peng Chen, Haobo Xu, Jiale Yao, Fang Li, and Tieyong Zeng. Learning cocoercive conservative denoisers via helmholtz decomposition for poisson imaging inverse problems. *Proc. NeurIPS*, 2025.

Lin Xiao, Stephen Boyd, and Sanjay Lall. A scheme for robust distributed sensor fusion based on average consensus. In *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.*, pp. 63–70. IEEE, 2005.

Masahiro Yukawa and Isao Yamada. Monotone Lipschitz-Gradient Denoiser: Explainability of Operator Regularization Approaches Free From Lipschitz Constant Control. *IEEE Trans. Signal Process.*, 73:3378–3393, 2025. ISSN 1053-587X, 1941-0476.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. *Proc. CVPR*, pp. 3929–3938, 2017.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10): 6360–6376, 2021.

APPENDIX

The appendix is organized as follows.

1. Appendix A presents the operator forms of the PnP and RED iterations.
2. Appendix B contains the proofs of the Metro-SSD properties and convergence guarantees.
3. Appendix C describes the efficient inverse-pair construction.
4. Appendix D provides the architecture, training, and warm-up details.
5. Appendix E reports denoising results, warm-up ablations, and additional visual examples.
6. Appendix F presents numerical convergence results.

A FIXED-POINT OPERATORS FOR PNP AND RED

We collect the fixed-point operators used in the convergence analysis here. For a denoiser $\mathcal{D}$, the PnP schemes considered in this work take the form $x_{k+1} = \mathcal{T}^{\mathcal{D}}(x_k)$, with

$$\begin{aligned}
&\text{(PnP-PGD)} \quad x_{k+1} = \mathcal{T}_{\gamma}^{\mathcal{D}}(x_k), \qquad \mathcal{T}_{\gamma}^{\mathcal{D}} := \mathcal{D} \circ (I - \gamma \nabla f),\\
&\text{(PnP-DRS)} \quad x_{k+1} = \mathcal{T}_{\mu,\beta}^{\mathcal{D}}(x_k), \qquad \mathcal{T}_{\mu,\beta}^{\mathcal{D}} := (1-\beta)I + \beta(2\,\mathrm{prox}_{\mu f} - I) \circ (2\mathcal{D} - I),\\
&\text{(PnP-HQS)} \quad x_{k+1} = \mathcal{T}_{\rho}^{\mathcal{D}}(x_k), \qquad \mathcal{T}_{\rho}^{\mathcal{D}} := \mathcal{D} \circ \mathrm{prox}_{\rho f}\,.
\end{aligned}$$

These correspond to PnP-PGD (Ryu et al., 2019; Hurault et al., 2022b), PnP-DRS (Ryu et al., 2019; Hurault et al., 2024), and PnP-HQS (Zhang et al., 2021; Hurault et al., 2022a), respectively. The parameters satisfy $\gamma, \rho, \mu > 0$ and $\beta \in (0, 1]$. For DRS, x_k denotes the splitting variable.

Similarly, RED-GD (Romano et al., 2017) uses

$$\text{(RED-GD)} \qquad x_{k+1} = \mathcal{T}_{\gamma,\lambda}^{\mathcal{D}}(x_k), \qquad \mathcal{T}_{\gamma,\lambda}^{\mathcal{D}} := I - \gamma\big(\nabla f + \lambda(I - \mathcal{D})\big),$$

where $\lambda > 0$.

For the data-fidelity term

$$f(x) = \frac{1}{2}\|Ax - y\|^2,$$

the gradient and proximal operator are

$$\nabla f(x) = A^\top Ax - A^\top y, \tag{15}$$

$$\operatorname{prox}_{\tau f}(v) = (I + \tau A^\top A)^{-1}(v + \tau A^\top y), \tag{16}$$

where $\tau > 0$. Introduce

$$B := A^\top A, \qquad c := A^\top y, \qquad R_\tau := (I + \tau B)^{-1}.$$

Since $B \succeq 0$, we have $I + \tau B \succ 0$, so R_τ is well defined even when A is rank deficient.

For a fixed linear denoiser $\mathcal{D}(x) = Wx$, substituting (15) and (16) into the operators above yields affine iterations of the form

$$x_{k+1} = \Phi x_k + \psi,$$

where Φ and ψ are fixed once the denoiser, forward operator, observations, and algorithm parameters are fixed. Using

$$(2\operatorname{prox}_{\mu f} - I)(v) = (2R_\mu - I)v + 2\mu R_\mu c$$

for PnP-DRS, the affine components are

$$\begin{aligned}
&\text{(PnP-PGD)} && \Phi_{\mathrm{PGD}} = W(I - \gamma B), && \psi_{\mathrm{PGD}} = \gamma W c,\\
&\text{(PnP-DRS)} && \Phi_{\mathrm{DRS}} = (1-\beta)I + \beta(2R_\mu - I)(2W - I), && \psi_{\mathrm{DRS}} = 2\beta\mu R_\mu c,\\
&\text{(PnP-HQS)} && \Phi_{\mathrm{HQS}} = WR_\rho, && \psi_{\mathrm{HQS}} = \rho W R_\rho c,\\
&\text{(RED-GD)} && \Phi_{\mathrm{RED}} = I - \gamma\big(B + \lambda(I - W)\big), && \psi_{\mathrm{RED}} = \gamma c.
\end{aligned} \tag{17}$$

For Metro-SSD, we set $W = W_{\mathrm{M}}$ in PnP-PGD, PnP-HQS, and RED-GD. For the contractive PnP-DRS construction, we use the averaged denoiser

$$W_{\mathrm{M},t} = (1-t)I + tW_{\mathrm{M}}, \qquad 0 < t \leqslant 1/2,$$

and set $W = W_{\mathrm{M},t}$ in (17). The PnP-DRS operator in (13) corresponds to $t = 1/2$, for which

$$2W_{\mathrm{M},1/2} - I = W_{\mathrm{M}}.$$

Thus, for quadratic data fidelity and a fixed Metro-SSD reference, the convergence analysis reduces to the linear part Φ of the corresponding affine iteration. In particular, if the spectral norm satisfies $\|\Phi\|_2 < 1$, then $I - \Phi$ is invertible and the iteration has the unique fixed point

$$x^* = (I - \Phi)^{-1}\psi, \qquad \|x_k - x^*\| \leqslant \|\Phi\|_2^k \|x_0 - x^*\|.$$

For complex-valued MRI operators, the same formulas hold with the transpose replaced by the Hermitian adjoint.

B PROOFS

Proof of Proposition 1. We first check that the Metropolis normalization is well defined. As $P_\pi e = e$ and $D_\pi e = \omega_\pi$, we have from (8) that

$$d = Ke = \sum_{\pi \in \mathcal{G}} \left(\omega_\pi + P_{\pi^{-1}}\omega_\pi\right).$$

Because $\mathcal{G}$ is nonempty and $\omega_\pi > 0$ for every $\pi \in \mathcal{G}$, it follows that $d > 0$ componentwise. Thus, for every $\pi \in \mathcal{G}$,

$$\max(d, P_\pi d) > 0,$$

and hence

$$M_\pi = \operatorname{Diag}\big(\max(d, P_\pi d)\big)$$

is invertible. Therefore K_{M} in (10), and hence W_{M} in (11), are well defined.

We next record the entrywise expressions that will be used below. By definition (6),

$$(P_\pi)_{ij} = \begin{cases} 1 & \text{if } i = \pi(j), \\ 0 & \text{otherwise.} \end{cases}$$

Hence, from (8), we have

$$K_{ij} = \sum_{\pi \in \mathcal{G}} \Big(\omega_\pi(i)(P_\pi)_{ij} + \omega_\pi(j)(P_{\pi^{-1}})_{ij} \Big) = \sum_{\substack{\pi \in \mathcal{G} \\ \pi(j) = i}} \omega_\pi(i) + \sum_{\substack{\pi \in \mathcal{G} \\ \pi(i) = j}} \omega_\pi(j). \tag{18}$$

A direct calculation in (10) using commutativity of the max operator gives

$$(K_{\mathrm{M}})_{ij} = \frac{K_{ij}}{\max(d_i, d_j)}.$$

We next establish symmetry, nonnegativity, and stochasticity of W_{M}. For every $\pi \in \mathcal{G}$, since D_π and M_π are diagonal and $P_\pi^\top = P_{\pi^{-1}}$,

$$\big(M_\pi^{-1} D_\pi P_\pi\big)^\top = P_{\pi^{-1}} D_\pi M_\pi^{-1}.$$

Thus K_{M} is symmetric. Moreover, D_π, M_π^{-1}, and P_π are nonnegative, so $K_{\mathrm{M}} \geqslant 0$.

To show that the diagonal correction in (11) is nonnegative, we first bound the row sums of K_{M}. For every i,

$$(K_{\mathrm{M}} e)_i = \sum_{j=1}^{n} \frac{K_{ij}}{\max(d_i, d_j)} \leqslant \frac{1}{d_i} \sum_{j=1}^{n} K_{ij} = 1.$$

It follows that, for every i,

$$(W_{\mathrm{M}})_{ii} = (K_{\mathrm{M}})_{ii} + 1 - (K_{\mathrm{M}} e)_i \geqslant 0,$$

while, for $i \neq j$,

$$(W_{\mathrm{M}})_{ij} = (K_{\mathrm{M}})_{ij} \geqslant 0.$$

Thus $W_{\mathrm{M}} \geqslant 0$. Since K_{M} is symmetric and the correction in (11) is diagonal, W_{M} is also symmetric. Finally,

$$W_{\mathrm{M}} e = K_{\mathrm{M}} e + e - K_{\mathrm{M}} e = e,$$

so W_{M} is stochastic.

It remains to prove primitivity under the additional assumptions that $\mathcal{G}$ is transitive and contains the identity permutation. We first show that K is irreducible. Fix any distinct $i, j \in \Omega$. By transitivity of $\mathcal{G}$, there exist $\pi_1, \dots, \pi_m \in \mathcal{G} \cup \mathcal{G}^{-1}$ such that

$$(\pi_m \circ \cdots \circ \pi_1)(i) = j.$$

Set $i_0 = i$ and

$$i_\ell = \pi_\ell(i_{\ell-1}), \qquad \ell = 1, \dots, m,$$

so that $i_m = j$. We claim that

$$K_{i_{\ell-1} i_\ell} > 0, \qquad \ell = 1, \dots, m.$$

Indeed, either $\pi_\ell \in \mathcal{G}$ or $\pi_\ell \in \mathcal{G}^{-1}$. If $\pi_\ell \in \mathcal{G}$, then $\pi_\ell(i_{\ell-1}) = i_\ell$, and (18) gives

$$K_{i_{\ell-1} i_\ell} \geqslant \omega_{\pi_\ell}(i_\ell) > 0.$$

On the other hand, if $\pi_\ell \in \mathcal{G}^{-1}$, then $\pi_\ell^{-1} \in \mathcal{G}$, so (18) gives

$$K_{i_{\ell-1} i_\ell} \geqslant \omega_{\pi_\ell^{-1}}(i_{\ell-1}) > 0.$$

Since K is nonnegative,

$$(K^m)_{ij} = \sum_{j_1, \dots, j_{m-1} \in \Omega} K_{i j_1} K_{j_1 j_2} \cdots K_{j_{m-1} j} \geqslant \prod_{\ell=1}^{m} K_{i_{\ell-1} i_\ell} > 0,$$

where the inequality follows by choosing $j_\ell = i_\ell$ for $\ell = 1, \dots, m-1$. Thus, for every distinct i, j, there exists $m \geqslant 1$ such that $(K^m)_{ij} > 0$. Hence K is irreducible.

By (9), since $d > 0$, for every $i \neq j$,

$$(K_{\mathrm{M}})_{ij} > 0 \quad \Longleftrightarrow \quad K_{ij} > 0.$$

Since (11) modifies only the diagonal, W_{M} has the same off-diagonal support and is therefore also irreducible.

Finally, let $\pi_{\mathrm{id}} \in \mathcal{G}$ denote the identity permutation. Since $P_{\pi_{\mathrm{id}}} = I$ and $\omega_{\pi_{\mathrm{id}}} > 0$, the corresponding term in (8) gives

$$K_{ii} \geqslant 2\,\omega_{\pi_{\mathrm{id}}}(i) > 0.$$

Consequently,

$$(W_{\mathrm{M}})_{ii} \geqslant (K_{\mathrm{M}})_{ii} = K_{ii}/d_i > 0.$$

Thus W_{M} is an irreducible nonnegative matrix with a positive diagonal, and is therefore primitive. □

Proof of Theorem 1. By assumption $Ae \neq 0$, so that

$$\mathrm{span}\{e\} \cap \mathrm{null}(A) = \{0\}. \tag{19}$$

Let

$$B = A^{\top}A, \qquad R_\tau = (I + \tau B)^{-1}, \qquad \tau > 0.$$

Throughout, $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced spectral norm for matrices and linear operators.

As shown in Appendix A, for quadratic data fidelity and a fixed Metro-SSD reference, the PnP-HQS and PnP-DRS operators are affine. Their contractivity is therefore determined entirely by the linear parts of the corresponding affine operators. Specifically, for PnP-HQS the linear part is

$$\Phi_{\mathrm{HQS}} = W_{\mathrm{M}} R_\rho.$$

For PnP-DRS, using the half-averaged denoiser $W_{\mathrm{M},1/2} = (I + W_{\mathrm{M}})/2$, whose reflected operator is W_{M}, the linear part becomes

$$\Phi_{\mathrm{DRS}} = (1-\beta)I + \beta(2R_\mu - I)W_{\mathrm{M}}.$$

We first collect some basic properties of the operators involved. Since W_{M} is symmetric, stochastic, and primitive, 1 is a simple eigenvalue with eigenvector e, and all remaining eigenvalues lie in $(-1, 1)$. By expanding x in an orthonormal eigenbasis of W_{M}, it follows that

$$\|W_{\mathrm{M}}x\| \leqslant \|x\|,$$

with equality if and only if $x \in \mathrm{span}\{e\}$.

We next study R_τ and its reflection $2R_\tau - I$. The matrix B is symmetric positive semidefinite and $\mathrm{null}(B) = \mathrm{null}(A)$. If $\lambda \geqslant 0$ is an eigenvalue of B, then the corresponding eigenvalues of R_τ and $2R_\tau - I$ are

$$\frac{1}{1+\tau\lambda} \qquad \text{and} \qquad \frac{1-\tau\lambda}{1+\tau\lambda}.$$

Since $\tau > 0$, the first lies in $(0, 1]$ and the second in $(-1, 1]$, with modulus one only when $\lambda = 0$. As before, expanding x in an orthonormal eigenbasis of B, it follows that

$$\|R_\tau x\| \leqslant \|x\|, \qquad \|(2R_\tau - I)x\| \leqslant \|x\|,$$

with equality in either case if and only if $x \in \mathrm{null}(B) = \mathrm{null}(A)$.

We first consider PnP-HQS. Its linear part is $\Phi_{\mathrm{HQS}} = W_{\mathrm{M}} R_\rho$. The preceding nonexpansiveness properties give

$$\|\Phi_{\mathrm{HQS}}\| = \|W_{\mathrm{M}} R_\rho\| \leqslant 1.$$

We claim that the inequality is in fact strict:

$$\|\Phi_{\mathrm{HQS}}\| < 1. \tag{20}$$

Suppose, to the contrary, that $\|\Phi_{\mathrm{HQS}}\| = 1$. Since the unit sphere in $\mathbb{R}^n$ is compact, there would exist $\|v\| = 1$ such that

$$\|\Phi_{\mathrm{HQS}} v\| = \|W_{\mathrm{M}} R_\rho v\| = 1.$$

But

$$\|W_{\mathrm{M}} R_\rho v\| \leqslant \|R_\rho v\| \leqslant \|v\| = 1.$$

Since the first and last quantities are equal, we have

$$\|W_{\mathrm{M}} R_\rho v\| = \|R_\rho v\|, \qquad \|R_\rho v\| = \|v\|.$$

The first equality implies $R_\rho v \in \mathrm{span}\{e\}$, while the second implies $v \in \mathrm{null}(A)$. Since $R_\rho v = v$ for $v \in \mathrm{null}(A)$, it follows from (19) that

$$v \in \mathrm{span}\{e\} \cap \mathrm{null}(A) = \{0\}.$$

This contradicts $\|v\| = 1$, establishing (20). Hence PnP-HQS is contractive for every $\rho > 0$.

We next consider PnP-DRS. Set $S_\mu := 2R_\mu - I$. From the properties established above,

$$\|S_\mu x\| \leqslant \|x\|,$$

with equality if and only if $x \in \mathrm{null}(A)$. The linear part of the PnP-DRS operator is

$$\Phi_{\mathrm{DRS}} = (1-\beta)I + \beta S_\mu W_{\mathrm{M}}.$$

Since both S_μ and W_{M} are nonexpansive, $\|S_\mu W_{\mathrm{M}}\| \leqslant 1$. We claim that the inequality is strict:

$$\|S_\mu W_{\mathrm{M}}\| < 1. \tag{21}$$

Suppose, to the contrary, that $\|S_\mu W_{\mathrm{M}}\| = 1$. Then $\|S_\mu W_{\mathrm{M}} v\| = 1$ for some $\|v\| = 1$. Since

$$\|S_\mu W_{\mathrm{M}} v\| \leqslant \|W_{\mathrm{M}} v\| \leqslant \|v\| = 1,$$

and the first and last quantities are equal, we have

$$\|S_\mu W_{\mathrm{M}} v\| = \|W_{\mathrm{M}} v\|, \qquad \|W_{\mathrm{M}} v\| = \|v\|.$$

The first equality implies $W_{\mathrm{M}} v \in \mathrm{null}(A)$, while the second implies $v \in \mathrm{span}\{e\}$. Since $W_{\mathrm{M}} v = v$ for $v \in \mathrm{span}\{e\}$, it follows from (19) that

$$v \in \mathrm{span}\{e\} \cap \mathrm{null}(A) = \{0\}.$$

This contradicts $\|v\| = 1$, establishing (21).

It follows that, for every $\beta \in (0, 1]$,

$$\|\Phi_{\mathrm{DRS}}\| = \|(1-\beta)I + \beta S_\mu W_{\mathrm{M}}\| \leqslant (1-\beta) + \beta \|S_\mu W_{\mathrm{M}}\| < 1.$$

Hence PnP-DRS is contractive for every $\mu > 0$ and $\beta \in (0, 1]$.

Thus, for both PnP-HQS and PnP-DRS, the corresponding linear part Φ satisfies $\|\Phi\| < 1$. Hence the full update operator $\mathcal{T}(x) = \Phi x + \psi$ is contractive. By the Banach contraction mapping theorem, $\mathcal{T}$ has a unique fixed point, and the iterates converge to it linearly from any initialization. □

C Inverse-Pair Implementation

For cyclic translations, the symmetric kernel in (8) can be efficiently implemented without predicting separate weight maps for a shift and its inverse (Sinha et al., 2026, Algorithm 1). Let $\mathcal{G}_{\mathrm{full}}$ denote the complete translation family and write

$$\mathcal{G}_{\mathrm{full}} = \{\pi_{\mathrm{id}}\} \,\dot\cup\, \mathcal{I}_+ \,\dot\cup\, \mathcal{I}_-, \qquad \mathcal{I}_- := \{\pi^{-1} : \pi \in \mathcal{I}_+\},$$

where $\mathcal{I}_+$ contains one representative from each non-identity inverse pair. For the local windows considered here, $2R < \min\{H, L\}$, so the identity is the only self-inverse translation.

The predictor is evaluated only on

$$\mathcal{G}_+ := \{\pi_{\mathrm{id}}\} \cup \mathcal{I}_+.$$

Thus, instead of $(2R+1)^2$ weight maps, it predicts

$$J = |\mathcal{G}_+| = 1 + \frac{|\mathcal{G}_{\mathrm{full}}| - 1}{2} = 2R^2 + 2R + 1.$$

For each $\pi \in \mathcal{I}_+$, let $D_\pi = \mathrm{Diag}(\omega_\pi)$. The contribution associated with π^{-1} is obtained from

$$(D_\pi P_\pi)^\top = P_{\pi^{-1}} D_\pi,$$

so no separate inverse weight map is needed. The identity term in (8) appears twice; since its weight is freely predicted, we absorb this fixed factor into the identity weight and write the resulting kernel as

$$K = D_{\pi_{\mathrm{id}}} + \sum_{\pi \in \mathcal{I}_+} \left(D_\pi P_\pi + P_{\pi^{-1}} D_\pi\right).$$

The same construction applies after Metropolis scaling. Let

$$d = Ke, \qquad M_\pi = \mathrm{Diag}\big(\max(d, P_\pi d)\big), \qquad \pi \in \mathcal{I}_+,$$

where the maximum is taken componentwise. Since $P_{\pi_{\mathrm{id}}} = I$, the Metropolis denominator for the identity term is simply d. Hence (10) can be implemented as

$$K_{\mathrm{M}} = \mathrm{Diag}(d)^{-1} D_{\pi_{\mathrm{id}}} + \sum_{\pi \in \mathcal{I}_+} \left(M_\pi^{-1} D_\pi P_\pi + P_{\pi^{-1}} D_\pi M_\pi^{-1} \right).$$

Thus, each predicted non-identity weight map is used for both directions of the corresponding inverse pair, with the same Metropolis denominator. Hence the inverse-pair construction avoids separate prediction and aggregation for inverse shifts while implementing the same symmetric Metropolis operator used in the analysis. The implementation requires two aggregation passes. The first computes Ke. The second computes $K_{\mathrm{M}}x$ and $K_{\mathrm{M}}e$ simultaneously, after which the diagonal correction in (11) gives $W_{\mathrm{M}}x$.

D Implementation Aspects

D.1 Architecture

The original NECTR implementation uses a convolutional feature extractor with pointwise residual blocks, followed by a pointwise network that scores each pair of reference and shifted features. This pairwise network is evaluated separately for every concatenated reference–shifted feature pair. Our predictor extracts the reference features once and projects them once into queries and keys. It then scores each shift using a slice of the projected key tensor and applies a softmax across them.

Feature extractor. The predictor in Sinha et al. (2026) infers pairwise weights directly from image intensities using a shallow network. We separate feature extraction from pairwise weight prediction so that the weights can depend on a wider spatial context. This change does not affect the Metropolis guarantees, which only require the predicted weights to be strictly positive. Specifically, a DRUNet-based encoder-decoder (Zhang et al., 2021), $\mathcal{F} : \mathbb{R}^{(C+1)\times H\times L} \mapsto \mathbb{R}^{F\times H\times L}$, receives the reference image z together with a spatially constant noise-level map and produces the full-resolution embedding

$$f = \mathcal{F}([z \mid \sigma e]) \in \mathbb{R}^{F\times H\times L}.$$

We apply LayerNorm across the channel dimension (Ba et al., 2016). The resulting F-dimensional vector at each spatial location represents the corresponding image node in its surrounding context. Thus, nodes with similar structures can receive similar representations even when their raw intensities differ. To compute attention over spatial shifts, we also form

$$f^{\mathrm{pad}} = \mathrm{pad}_R(f) \in \mathbb{R}^{F\times(H+2R)\times(L+2R)}.$$

We project the unpadded embedding into central queries and the circularly padded embedding into keys. Both embeddings come from a single evaluation of the feature extractor.

Pairwise multi-head attention. We learn pairwise affinities from the contextual embeddings instead of prescribing a similarity function based on raw pixel or patch differences. Multiple heads allow the model to compare neighboring nodes in complementary feature subspaces (Vaswani et al., 2017; Wang et al., 2018). Two learned 1×1 convolutions project the padded and unpadded embeddings into one query tensor and one key tensor, respectively. Each tensor contains m heads of dimension d:

$$\mathcal{Q} = \mathrm{reshape}_{m,d}(\mathcal{V}_{\mathcal{Q}} * f^{\mathrm{pad}}) \in \mathbb{R}^{m\times d\times(H+2R)\times(L+2R)},$$
$$\mathcal{K} = \mathrm{reshape}_{m,d}(\mathcal{V}_{\mathcal{K}} * f) \in \mathbb{R}^{m\times d\times H\times L}.$$

We apply LayerNorm independently to each query and key vector within each head (Henry et al., 2020). For each $\pi \in \mathcal{G}_+$, we extract the $H \times L$ query window at the corresponding spatial offset and score it against the central key:

$$\mathcal{Q}_\pi = \mathrm{slice}_\pi(\mathcal{Q}) \in \mathbb{R}^{m\times d\times H\times L}, \qquad s_\pi = \frac{1}{\sqrt{d}} \langle \mathcal{Q}_\pi, \mathcal{K} \rangle_d \in \mathbb{R}^{m\times H\times L}.$$

The implementation computes these scores one shift at a time, without forming a tensor of all shifted query or key vectors. It then collects the score maps and normalizes them across shifts:

$$s = \mathrm{stack}_{\pi\in\mathcal{G}_+}(s_\pi) \in \mathbb{R}^{J\times m\times H\times L}, \qquad \widetilde{\omega} = \mathrm{softmax}_J(s) \in \mathbb{R}^{J\times m\times H\times L}_{++}.$$

The inner product is taken along the d-dimensional feature axis. The softmax acts across shifts independently at each head and spatial location, producing strictly positive affinities.

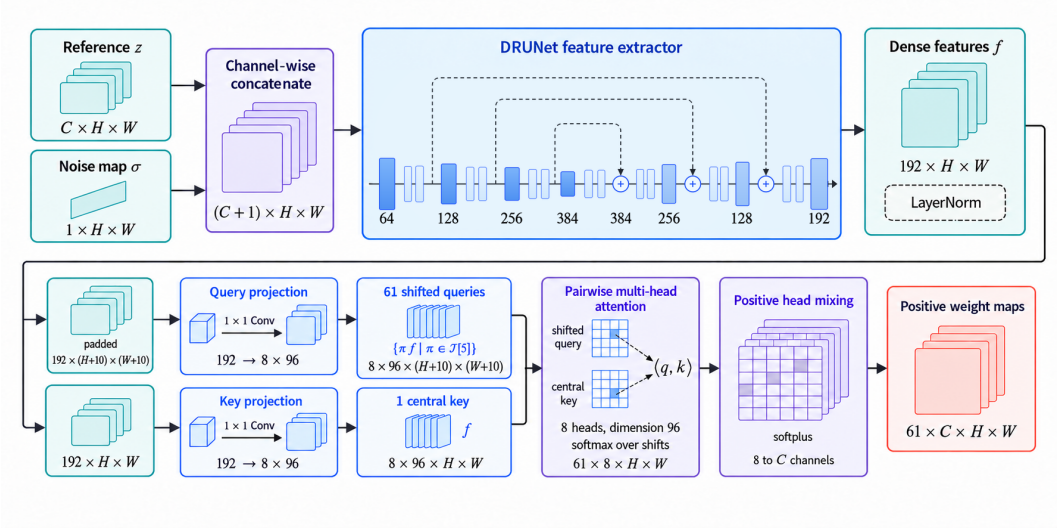


Figure 4: Architecture of the weight predictor $\mathcal{N}_\theta$. The reference image z is first concatenated with a spatially constant noise-level map and processed by the backbone to produce full-resolution features. The attention module projects the features once, scores successive shifts using slices of the padded key tensor, and normalizes the scores across shifts. A head-mixing layer combines the attention outputs to produce channel-specific weight maps. The softmax normalization of the affinities, together with softplus-parameterized mixing coefficients, ensures that all predicted weights are strictly positive.

Head mixing. We combine the eight head-wise affinity maps into the three image channels using a learned positive map $\mathcal{V}_O \in \mathbb{R}^{C \times m}$:

$$\omega_{\mathcal{G}} = \widetilde{\omega} \times_{\text{head}} \text{softplus}(\mathcal{V}_O) \in \mathbb{R}_{++}^{J \times C \times H \times L}.$$

Here, $\times_{\text{head}}$ denotes multiplication along the head axis, and the slice indexed by $\pi \in \mathcal{G}$ gives ω_π. Both the softmax affinities and the softplus-parameterized mixing coefficients are strictly positive. Consequently, every predicted weight satisfies the positivity condition of the Metropolis construction while remaining channel specific and data adaptive.

Memory usage. Our implementation does not materialize a $J \times m \times d \times H \times L$ tensor of shifted features or projections. It computes the query and padded key projections once and takes a key slice for each shift. The resulting score maps are then collected in a $J \times m \times H \times L$ tensor for normalization. These scores and the predicted weight maps still require memory that grows with the number of shifts and the image size. We emphasize that here, *matrix-free* means that the full $n \times n$ denoising matrix, with $n = CHL$, is never explicitly constructed. For example, a 256×256 RGB image has $n = 196{,}608$, so that dense matrix would contain approximately 3.87×10^{10} entries. Our operator-level implementation avoids this prohibitive representation by exploiting the structured shift interactions and computing the matrix action directly through elementwise multiplication and division, and reductions over the shifts.

D.2 Training Details

We train the proposed Metropolis denoiser W_{M} as a single-step Gaussian denoiser by minimizing the empirical mean squared error (MSE) with the Adam optimizer (Kingma & Ba, 2017):

$$\mathcal{L}(\theta) = \mathbb{E}_{\bar{x},\sigma,\epsilon}\left[\left\|W_{\text{M}}^{\bar{x}+\epsilon}(\bar{x}+\epsilon) - \bar{x}\right\|_2^2\right].$$

Here, the noisy patch $z = \bar{x} + \epsilon$ serves as both the denoiser input and the reference image. Following (Zhang et al., 2021), we randomly sample 8 clean patches, of size 128×128, $\bar{x} \sim \mathsf{D}$ per iteration. The datset D comprises of BSD400 (Martin et al., 2001), DIV2K (Timofte et al., 2017), the Waterloo Exploration Database (Ma et al., 2017), and Flickr2K (Lim et al., 2017) We corrupt each patch with additive white Gaussian noise, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma \sim \mathcal{U}[0, 50/255]$. A cosine schedule decreases the learning rate from 10^{-4} to 5×10^{-7} over $510{,}000$ iterations. We validate on BSD32 and checkpoint every $30{,}000$ iterations.

D.3 Warm-up Schemes

The warm-up procedure described in the main text uses the outer reconstruction iterate as the reference. Our implementation instead collects the image entering the denoiser at each warm-up iteration. These images differ from the outer iterates because of the intermediate data-fidelity and splitting updates. We observed better reconstruction quality with this procedure in our experiments.

PnP-HQS. Starting from x_0, the warm-up updates are

$$z_k = \mathrm{prox}_{\rho_{\mathrm{w}} f}(x_k), \qquad x_{k+1} = W_{\mathrm{M}}^{z_k}(z_k), \qquad k = 0, \dots, Q-1, \tag{22}$$

where $\rho_{\mathrm{w}} > 0$ is the warm-up proximal parameter. We retain the proximal output z_k, rather than x_k or the denoised output x_{k+1}, as the candidate reference.

PnP-DRS implementation. For MRI, we use the denoiser-first implementation (13) associated with the DRS splitting Wei et al. (2025). Its state consists of a denoised image u_k, a data-fidelity variable v_k, and a scaled dual variable b_k. Starting from $u_0 = v_0 = x_0$ and $b_0 = 0$, the updates are

$$\begin{aligned} z_k &= v_k - b_k, \\ u_{k+1} &= 0.5(I + W_{\mathrm{M}}^{z_k})(z_k), \\ a_{k+1} &= 2\beta u_{k+1} + (1-2\beta)v_k, \qquad k = 0, \dots, Q-1, \\ v_{k+1} &= \mathrm{prox}_{\mu_{\mathrm{w}} f}(a_{k+1} + b_k), \\ b_{k+1} &= b_k + a_{k+1} - v_{k+1}, \end{aligned} \tag{23}$$

where $\mu_{\mathrm{w}} > 0$ and β controls relaxation. For $\beta = 1/2$, $a_{k+1} = u_{k+1}$, giving the standard denoiser-first ADMM updates. The candidate reference is $z_k = v_k - b_k$, the input to the denoiser, whereas u_{k+1} is the reconstruction returned by the implementation. These variables should not be confused with the single splitting variable in the operator formulation of DRS (Ryu et al., 2019; Hurault et al., 2022b; 2024).

Reference freezing. For both algorithms, we use the guide candidate produced by the final warm-up iteration as the fixed reference, and denote it by z_Q. This is done since selecting the highest-PSNR candidate uses information unavailable to a ground-truth-free procedure. After warm-up, we hold z_Q fixed and use $W_{\mathrm{M}}^{z_Q}$ for PnP-HQS and $0.5(I + W_{\mathrm{M}}^{z_Q})$ for PnP-DRS. The iterations continue from their current states without restarting. At this point, we can also switch from the warm-up parameters to the prescribed reconstruction step size and denoising noise level. The convergence analysis applies to this subsequent phase, during which the reference and algorithm parameters remain fixed.

E Additional Results and Ablations

E.1 Denoising Performance

We first evaluate the standalone denoising performance of the directly trained Metropolis operator W_{M} against established baselines on the CBSD68 (Martin et al., 2001) dataset. These include classical filters such as BM3D (Dabov et al., 2007a) and CBM3D (Dabov et al., 2007b), as well as trained neural networks such as DRUNet (Zhang et al., 2021), GSDRUNet (Hurault et al., 2022a), ProxDRUNet (Hurault et al., 2022b), DEAL (Pourya et al., 2025) and the NECTR (Sinha et al., 2026) baseline. The peak signal-to-noise ratio (PSNR) results are reported in Table 4. Because the same operator is used during training and inference, the proposed Metropolis normalization avoids the performance loss associated with post-hoc symmetrization. Visual comparisons are provided in Figures 5 and 6.

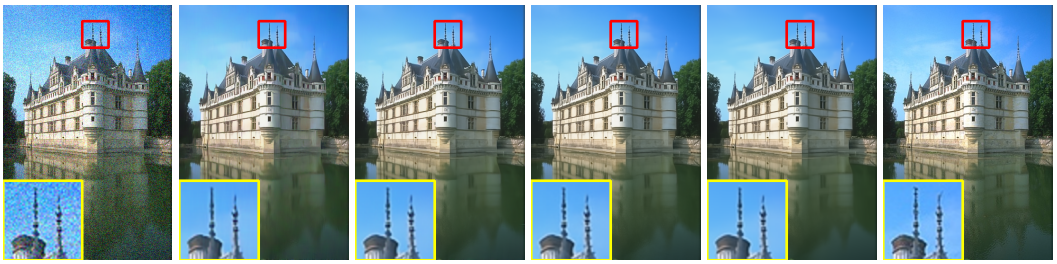
Summary. On CBSD68, Metro-SSD improves on the NECTR baseline across the reported noise levels and approaches the performance of the unconstrained DRUNet denoiser. The examples in Figures 5 and 6 show comparable preservation of image structure at two noise levels.

E.2 Warm-up Schemes

We study how reconstruction quality depends on the warm-up length and on the image supplied to the denoiser as its reference. The outer-iterate scheme in (14) uses the current reconstruction, whereas the inner schemes in (22)–(23) uses the input to the denoiser.

Table 4: Denoising performance on CBSD68 across noise levels $\nu/255$.

Method	$\nu = 5$	$\nu = 15$	$\nu = 25$	$\nu = 50$	Avg
Noisy	34.15	24.61	20.17	14.15	22.55
BM3D (Dabov et al., 2007a)	37.52	31.18	28.68	25.77	30.79
CBM3D (Dabov et al., 2007b)	—	33.44	30.69	27.35	—
DRUNet (Zhang et al., 2021)	40.65	34.30	31.68	28.49	33.78
GSDRUNet (Hurault et al., 2022a)	40.57	34.24	31.61	28.41	33.71
ProxDRUNet (Hurault et al., 2022b)	40.41	33.90	31.13	27.58	33.26
SPCDRUNet (Wei et al., 2024)	40.38	34.04	31.38	28.09	33.47
CoCoDRUNet (Wei et al., 2025)	40.10	33.82	31.15	27.86	33.23
DEAL (Pourya et al., 2025)	40.31	33.93	31.28	28.07	33.40
NECTR (Sinha et al., 2026)	39.60	33.52	30.86	27.53	32.88
Metro-SSD	40.24	34.11	31.52	28.31	33.55



noisy 20.52 dB | NECTR 30.91 dB | Metro-SSD 31.50 dB | GSDRUNet 31.68 dB | DRUNet 31.74 dB | clean

Figure 5: Denoising results at noise level $25/255$.

As shown in Table 5, even a short warm-up substantially improves on fixing the reference at initialization. For deblurring and superresolution, the gains become small by $Q = 30$; the MRI score continues to rise slightly beyond $Q = 150$. We use $Q = 30$ and $Q = 150$, respectively, as these settings capture most of the observed improvement without extending warm-up further.

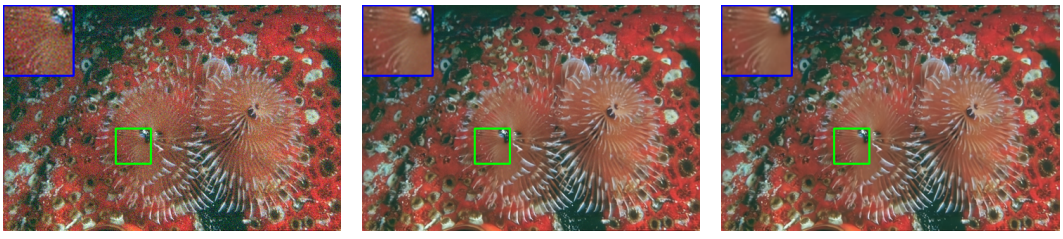
At the same warm-up length, the inner scheme gives higher PSNR in all four settings in Table 6. The difference is largest for MRI, at 2.63 dB. Warm-up performance therefore depends both on how long the reference is adapted and on which intermediate image is used to form the fixed linear denoiser.

Summary. Warm-up substantially improves reconstruction quality over fixing the reference at initialization. The gains largely level off by $Q = 30$ for deblurring and superresolution, while MRI benefits from a longer warm-up. At the chosen lengths, using the denoiser input as the reference candidate gives higher PSNR than using the outer iterate in all four settings. Thus, both the warm-up length and the image used to form the fixed denoiser matter.

E.3 Visual Comparisons

We provide additional visual results for motion deblurring, Gaussian deblurring, superresolution, and MRI. In the motion-deblurring examples in Figures 7 and 8, Metro-SSD recovers the repeated structures of the bridge and the architectural edges of the castle, with slightly higher PSNR than the displayed baselines. The Gaussian-deblurring example in Figure 9 shows comparable recovery of the dome's framework, although fine structures remain less distinct than in the clean image for all methods. For superresolution, Figures 10 and 11 consider challenging scenes containing dense and irregular textures. Metro-SSD's reconstruction remain visually competitive, with comparable recovery of the prominent structures and fine textures.

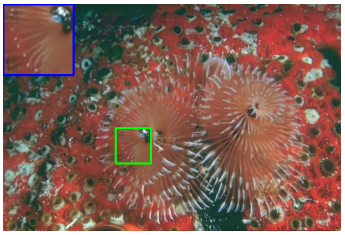
The MRI examples in Figures 12 and 13 cover both single- and multi-coil acquisitions. Metro-SSD reduces the artifacts present in the zero-filled reconstructions while retaining the anatomical boundaries visible in the highlighted regions. The competing methods produce broadly similar reconstructions, with Metro-SSD attaining modestly higher PSNR in these examples. These comparisons support the quantitative results across structures or acquisition settings.



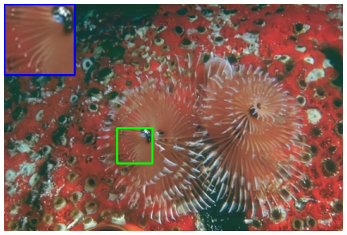
noisy
24.65 dB

ProxDRUNet
32.81 dB

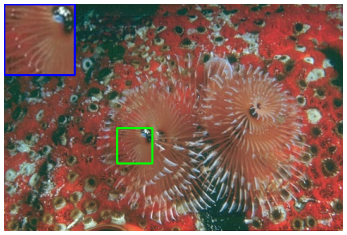
Metro-SSD
32.83 dB



GSDRUNet
32.92 dB



DRUNet
33.04 dB



clean

Figure 6: Denoising results at noise level $15/255$.

Table 5: Effect of warm-up length on the mean final PSNR (dB). "Start" reports the PSNR of the initialization, while 0 denotes the case in which the reference image is fixed from initialization and no warm-up is performed. We evaluate warm-up lengths from $5t$ to $50t$, with $t = 1$ for deblurring and superresolution and $t = 5$ for MRI. The highlighted $30t$ column corresponds to the warm-up length used in the main experiments. For deblurring and superresolution, the noise level is 0.03; for MRI, it is 0.002, with results averaged over the first ten images. All experiments use the inner warm-up scheme described in Appendix D.3.

Application	Start	0	$5t$	$10t$	$20t$	$30t$	$40t$	$50t$
Motion deblurring, kernel 1	21.48	24.82	28.30	28.95	29.20	29.23	29.22	29.22
Gaussian deblurring	23.48	26.35	27.62	27.72	27.75	27.78	27.77	27.76
$2\times$ superresolution, kernel 2	24.08	26.45	27.45	27.79	27.80	27.80	27.81	27.81
2-fold single-coil PD MRI	36.22	40.61	41.96	42.84	43.71	44.05	44.10	44.11

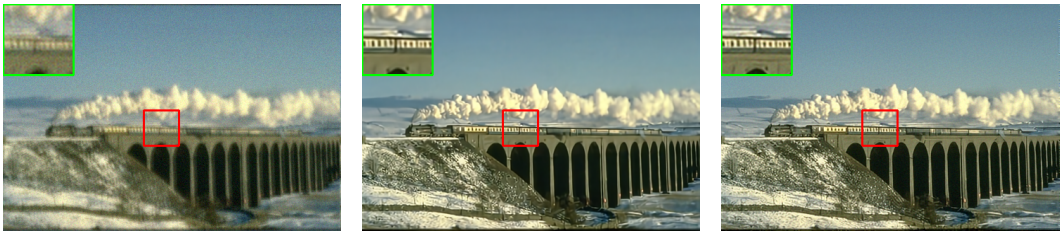
F Convergence Diagnostics

Figure 14 illustrates the convergence behavior of Metro-SSD with different PnP algorithms across deblurring, superresolution, and MRI. We use PnP-HQS (12) for deblurring with kernel 6 and noise level $\nu = 0.04$, and for $2\times$ superresolution with kernel 1 and $\nu = 0.02$. For MRI, we use PnP-DRS (13) on eight-fold accelerated multi-coil PDFS images from the dataset of Goujon et al. (2023). Here, Metro-SSD is half-averaged with the identity, i.e., $0.5(I + W_{\mathrm{M}})$, to retain the convergence guarantee in Theorem 1. We use a warm-up length of $Q = 30$ for deblurring and superresolution and $Q = 150$ for MRI. The abrupt transient near the corresponding warm-up boundary occurs because the reference image evolves during warm-up and is then frozen at $z = x_Q$, producing a one-time change in the reconstruction dynamics. This should not be interpreted as instability: the contractivity guarantee applies to the subsequent fixed-reference phase, in which the PSNR approaches a stable value and the normalized successive-iterate difference, $\|x_k - x_{k-1}\|/\sqrt{n}$ with $n = C \times H \times L$, decays toward zero. The bold curves show the dataset means, while the shaded regions indicate one standard deviation.

We further illustrate the practical consequence of contractivity in Figure 15. Across deblurring, superresolution, and MRI, the PnP iterations initialized from widely different points converge to the same reconstruction, as reflected by their PSNR trajectories approaching a common value. This empirically demonstrates that, for a fixed reference image, PnP-HQS and PnP-DRS with Metro-SSD converge to a unique fixed point independently of initialization.

Table 6: Effect of the warm-up scheme on the mean final PSNR (dB) for the tasks in Table 5. Both schemes use $Q = 30t$ warm-up iterations, with $t = 1$ for deblurring and superresolution and $t = 5$ for MRI. The "Inner" scheme uses the denoiser input as the candidate reference, as described in Appendix D.3, whereas the "Outer" scheme uses the current reconstruction, as in Section 4.1. The highlighted entries correspond to the inner warm-up scheme.

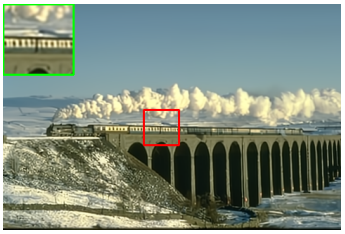
Scheme	Motion deblur	Gaussian deblur	2× SR	2-fold PD MRI
Inner	29.23	27.78	27.80	44.05
Outer	28.42	27.67	27.62	41.42



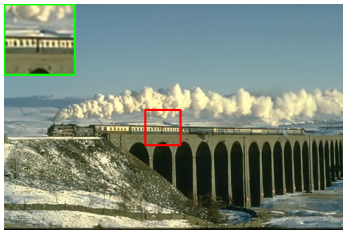
blurry
20.45 dB

ProxPnP-PGD
27.57 dB

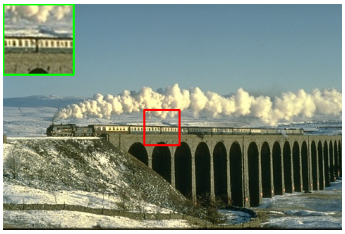
GSPnP-HQS
27.50 dB



DPIR
27.43 dB



Metro-SSD
27.72 dB



clean

Figure 7: Motion deblurring results on "train" image using kernel 5 at noise level 0.03.

Summary. After warm-up, the PSNR and successive-iterate curves stabilize across the three tasks. The trajectories from different initializations also approach the same reconstruction when the reference is fixed, consistent with the fixed-point convergence guarantees.

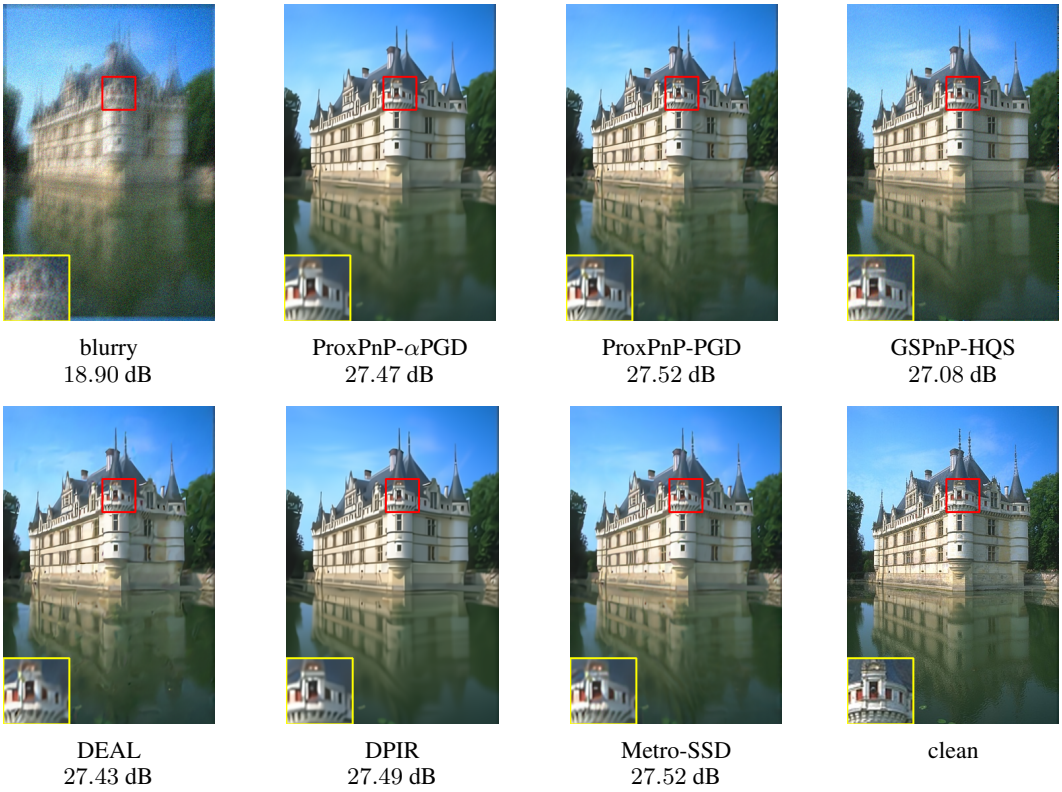


Figure 8: Motion deblurring results on "castle" image using kernel 8 at noise level 0.05.

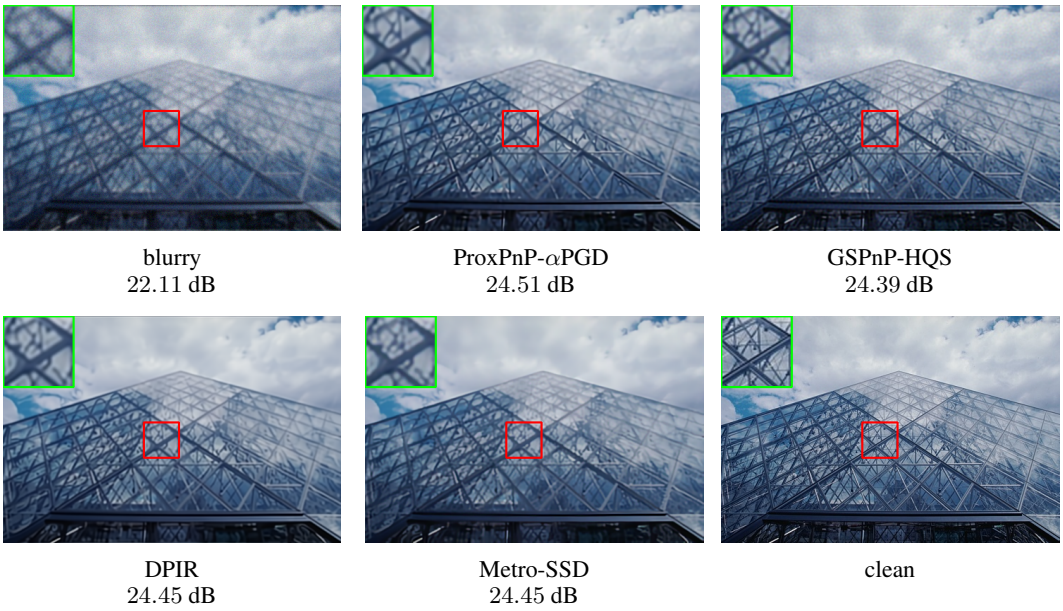
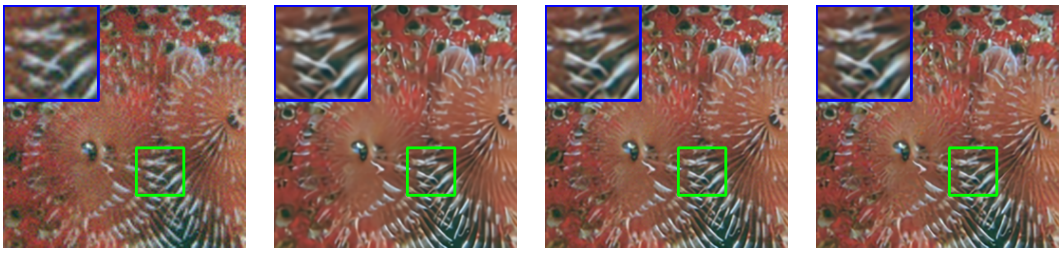


Figure 9: Deblurring of "dome" using a 25×25 Gaussian kernel with standard deviation 1.6, and at noise level 0.05.

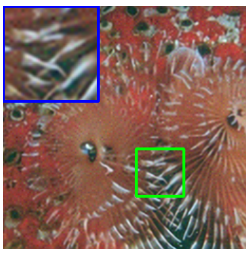


bicubic
25.37 dB

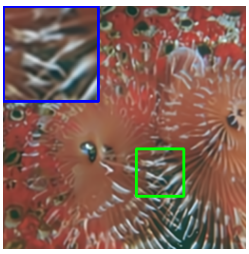
ProxPnP-αPGD
29.21 dB

ProxPnP-DRS
29.08 dB

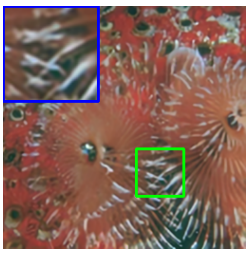
DEAL
29.04 dB



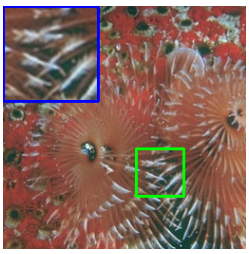
GSPnP-HQS
28.40 dB



DPIR
29.27 dB



Metro-SSD
29.11 dB



clean

Figure 10: $2\times$ superresolution of "coral" image with kernel 1 and noise level $\nu = 0.03$.



bicubic
19.83 dB

ProxPnP-αPGD
22.70 dB

ProxPnP-PGD
22.54 dB

DEAL
22.57 dB



GSPnP-HQS
23.22 dB



DPIR
23.06 dB

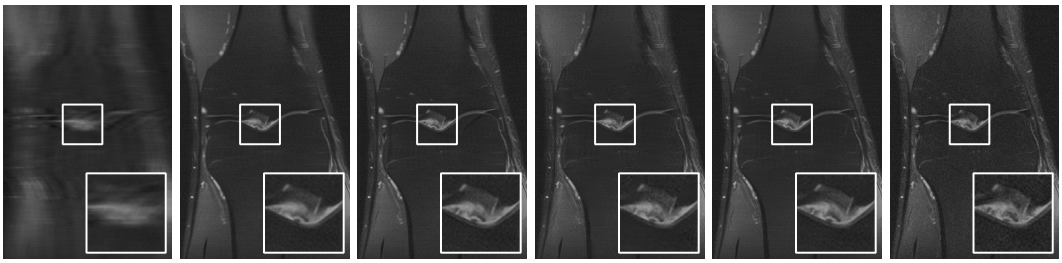


Metro-SSD
23.01 dB



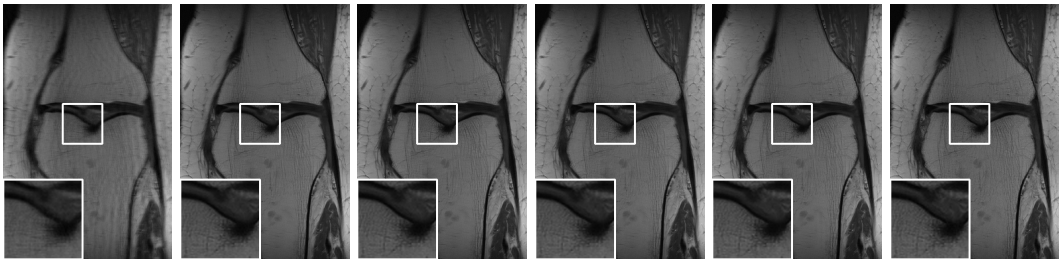
clean

Figure 11: $2\times$ superresolution of "clock-tower" image with kernel 4 and noise level $\nu = 0.03$.



zero-fill $(A^\top y)$ 30.55 dB | ProxPnP-DRS 35.70 dB | GSPnP-HQS 36.27 dB | DEAL 35.39 dB | Metro-SSD 36.33 dB | clean

Figure 12: Knee-MRI reconstruction for the 8-fold multicoil PDFS setting with and noise level 0.002.



zero-fill $(A^\top y)$ 34.78 dB | ProxPnP-DRS 44.58 dB | GSPnP-HQS 44.32 dB | DEAL 44.55 dB | Metro-SSD 44.72 dB | clean

Figure 13: Knee-MRI reconstruction for 2-fold singlecoil PD with noise level 0.002.

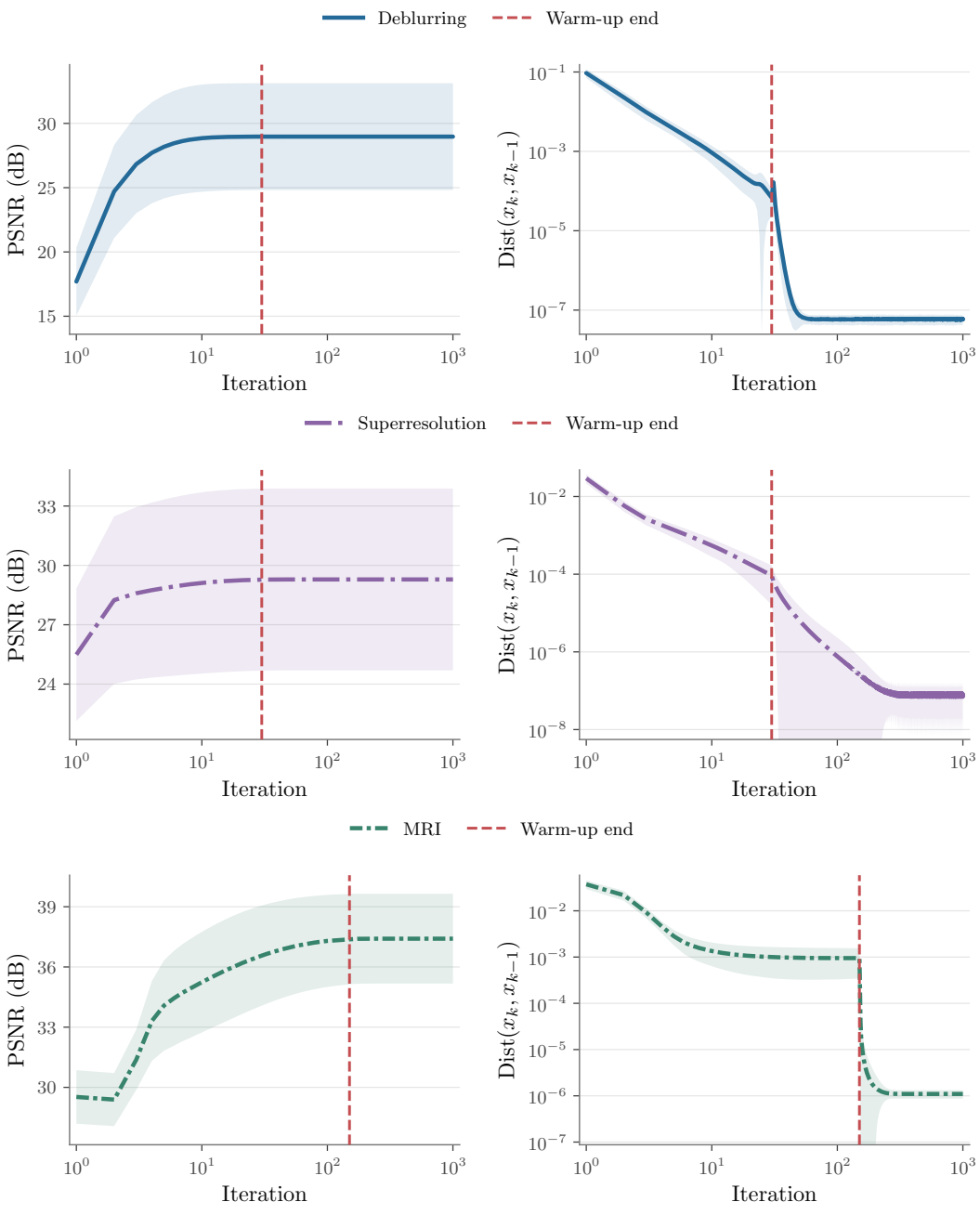
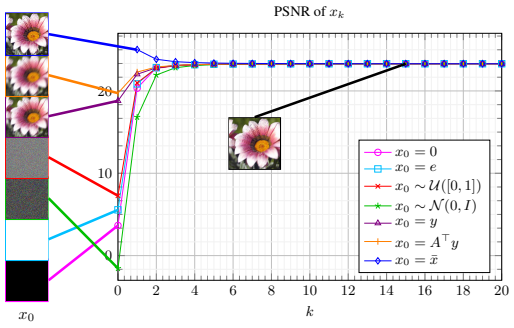
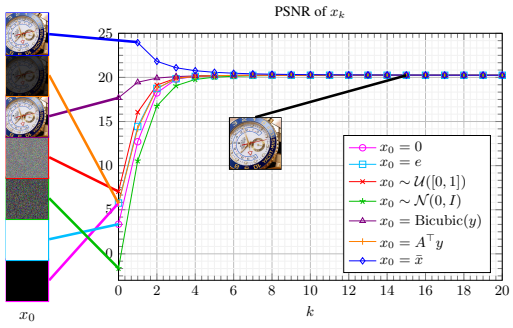


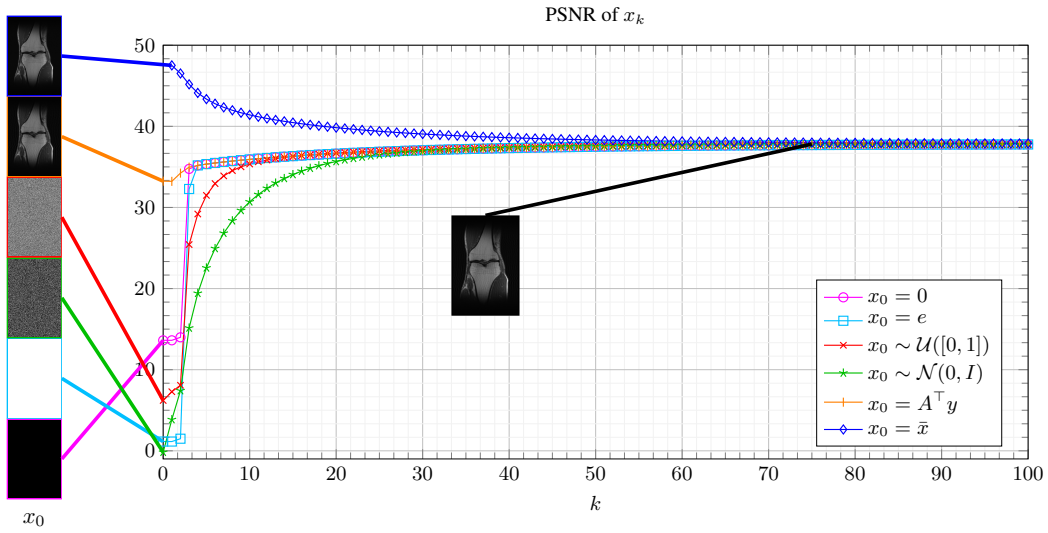
Figure 14: **Convergence diagnostics for Metro-SSD.** The top, middle, and bottom rows are the results for PnP-HQS deblurring with kernel 6 and $\nu = 0.04$, $2\times$ superresolution using PnP-HQS with kernel 1 and $\nu = 0.02$, and eight-fold accelerated multi-coil PDFS MRI using PnP-DRS. Results are averaged over the corresponding datasets. For MRI, Metro-SSD is half-averaged with the identity, as required by Theorem 1. We use warm-up lengths of $Q = 30$ for deblurring and superresolution and $Q = 150$ for MRI. The brief transient at the warm-up boundary occurs when the evolving reference image is frozen, causing a one-time change in the reconstruction operator. After this transition, the PSNR stabilizes and the normalized successive-iterate difference, decays toward zero. Bold curves show the dataset mean, while shaded regions indicate one standard deviation.



(a) Deblurring



(b) Superresolution



(c) MRI

Figure 15: **Consequence of contractivity.** Motion deblurring and $2\times$ superresolution are shown for *lily* and *watch* images from General100 (Dong et al., 2016), using kernels 8 and 1, respectively, at noise level $\nu = 0.03$. The MRI example uses a two-fold accelerated multi-coil PD image from the dataset of (Goujon et al., 2023). For each task, different initializations converge to the same reconstruction. These trajectories are consistent with the contractivity of the corresponding PnP operators with Metro-SSD (11) and a fixed reference image, which guarantees convergence to a unique fixed point independently of the initialization x_0.