LoTA-N2N: Local Trace Adaptation for Zero-Shot Self-Supervised Image Denoising

Anonymous authors
Paper under double-blind review

Abstract

Many self-supervised image denoising methods minimize the mean squared error (MSE) between denoiser predictions and surrogate targets constructed from noisy observations. They justify this training under different assumptions, such as suitable noise independence and unbiasedness, weak corruption, clean-neighbor similarity, or a compatible corruption model. We revisit these training objectives and show that self-supervised MSE differs from supervised MSE by a parameter-independent constant and a trace term. From this perspective, existing assumptions eliminate or suppress the trace term, narrowing the gap between self-supervised and supervised objectives up to a constant. This helps explain why training with surrogate targets generated from noisy images can yield effective denoising. Motivated by this understanding, we propose to estimate and penalize its magnitude alongside reconstruction, reducing reliance on assumptions that suppress it implicitly. However, a small global trace can conceal large regional discrepancies through cancellation of positive and negative contributions. We therefore introduce LoTA-N2N, a two-stage zero-shot denoising framework with local trace correction. A denoiser first learns from complementary sub-images of a single noisy observation and is frozen to provide clean-image proxies. These proxies enable interaction estimates whose absolute magnitudes are penalized region by region during adaptation. We prove that this local penalty prevents cancellation between regions and yields a supervised-risk upper bound with slack controlled by teacher-estimation error. Experiments demonstrate improvements over iteration-matched MSE adaptation under independent, spatially varying, and mixed noise.

1 Introduction

Deep image denoisers are commonly trained by minimizing the discrepancy between their predictions and clean reference images. Acquiring such references, however, is often difficult in microscopy, medical imaging, and real camera pipelines, where repeated acquisition can introduce motion, photobleaching, or registration errors Zhang et al. (2017; 2018); Anwar & Barnes (2019); Guo et al. (2019). Self-supervised denoising addresses this difficulty by constructing substitute targets from noisy observations. This replacement makes training possible without clean references, but also changes the objective being optimized. The central question is therefore *when, and why, minimizing an observable self-supervised loss also reduces the reconstruction error with respect to the unknown clean image*.

Existing methods answer this question under different assumptions about the observations and the construction of training pairs. Noise2Noise and blind-spot methods establish the validity of noisy targets through suitable conditional independence and unbiasedness conditions Lehtinen et al. (2018); Krull et al. (2019); Batson & Royer (2019); Laine et al. (2019). Noisy As Clean uses weakly corrupted observations as approximate clean targets Xu et al. (2020), while re-corruption methods construct training pairs using knowledge of the corruption process Pang et al. (2021); Monroy et al. (2025). Neighbor2Neighbor and ZSN2N exploit similarity between neighboring clean pixels together with suitable noise conditions Huang et al. (2021); Mansour & Heckel (2023). Each approach has its own justification for why its assumptions enable denoising. We revisit these justifications through a common question: **how do these assumptions affect the difference between the self-supervised objective and its supervised counterpart?**

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.

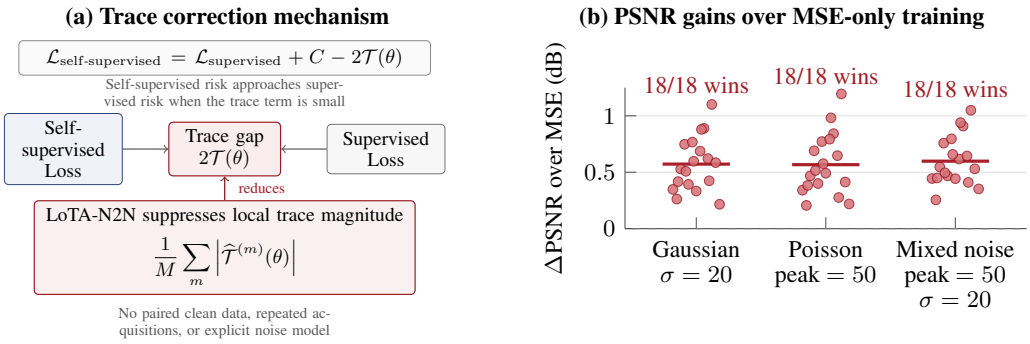


Figure 1: **Improving denoising with local trace correction.** (a) The trace term captures the parameter-dependent difference between self-supervised and supervised MSE objectives. LoTA-N2N estimates this interaction and penalizes its magnitude within each region to prevent spatial cancellation. (b) Per-image PSNR gains over MSE-only adaptation on McMaster18 under Gaussian, Poisson, and mixed noise. Each point represents one image, and horizontal bars indicate mean gains

For a broad class of MSE-based formulations, we characterize this difference through a single residual–prediction interaction. Writing the substitute target as the clean target plus a residual gives

$$\mathcal{L}_{\text{ss}}(\theta) = \mathcal{L}_{\text{sup}}(\theta) + C - 2\mathcal{T}(\theta),$$

where C is parameter-independent and $\mathcal{T}(\theta)$ measures the interaction between the target residual and the prediction error. This trace term provides a common interpretation of the assumptions above: conditional independence and unbiasedness can eliminate the interaction, weak noise and clean-neighbor similarity can limit its magnitude, and a compatible re-corruption model can ensure its cancellation. These assumptions thus enable self-supervised learning by eliminating or limiting the parameter-dependent discrepancy from supervised MSE.

This interpretation suggests a direct design principle: estimate and penalize the discrepancy that existing assumptions are intended to suppress. Instead of requiring the trace to vanish through the training construction, we constrain its magnitude alongside reconstruction MSE. This offers a way to reduce reliance on exact independence, negligible target mismatch, or a prescribed corruption model, provided that the remaining interaction can be estimated reliably.

However, a small global trace does not ensure small discrepancies within individual regions. Because the trace is signed, regional interactions of $+a$ and $-a$ cancel globally despite their potentially large magnitudes. We therefore penalize the absolute trace within each region before averaging. This local construction prevents cancellation between regions and upper-bounds the absolute global trace, making regional objective mismatch an explicit training penalty.

We implement this principle in LoTA-N2N, a two-stage zero-shot denoising framework. A denoiser first learns from complementary sub-images of one noisy observation and is frozen to provide clean-image proxies. These proxies enable local interaction estimates, which guide further adaptation through bidirectional MSE and local trace correction (LTrC). We derive a supervised-risk upper bound with slack controlled by teacher-estimation error. Experiments on natural, confocal, and X-ray images show gains over iteration-matched MSE adaptation under independent, spatially varying, and mixed noise, while correlated-noise tests expose the limits of the sub-image construction.

Our main contributions are as follows:

- We provide a trace-based interpretation of representative MSE self-supervised denoising methods, showing how their assumptions eliminate or limit a common objective discrepancy and motivating its explicit control.
- We propose LoTA-N2N, which estimates this discrepancy using a frozen single-image teacher and penalizes its local absolute magnitude. We establish cancellation control and a supervised-risk upper bound that accounts for teacher-estimation error.
- We demonstrate improvements over iteration-matched MSE adaptation across multiple image domains and noise settings, and use controlled ablations to examine the benefits of local control and the limits of the sub-image construction.

2 Related Work

Paired and blind-spot supervision. Noise2Noise learns from conditionally independent, unbiased noisy observations of the same scene Lehtinen et al. (2018). Blind-spot methods instead predict held-out pixels from their context, preventing direct access to the target noise Krull et al. (2019); Batson & Royer (2019); Laine et al. (2019). For real images, spatially correlated noise can still enter through neighboring pixels. AP-BSN uses asymmetric pixel-shuffle downsampling Lee et al. (2022), while MASH combines masking with local shuffling Chihaoui & Favaro (2024). BSN-INS uses implicit neural estimates to complement distant visible pixels Quan et al. (2025), and TM-BSN designs blind regions around the spatial structure of noise correlation Park et al. (2026). These approaches seek to limit target-noise leakage while preserving useful image context.

Weak-noise and re-corruption methods. Noisy As Clean treats a weakly corrupted observation as an approximate clean target Xu et al. (2020). Noisier2Noise, R2R, and IDR construct additional noisy inputs or refined targets using assumptions about the corruption process Moran et al. (2020); Pang et al. (2021); Zhang et al. (2022). Generalized R2R extends this approach beyond Gaussian noise, including Poisson and Gamma distributions, and establishes a connection to supervised squared loss under suitable noise conditions Monroy et al. (2025). These methods illustrate how a compatible corruption model aligns noisy-target reconstruction with the supervised objective.

Image self-similarity and single-image training. Neighbor2Neighbor constructs training pairs from neighboring pixels Huang et al. (2021), while ZSN2N and Noise2Fast enable efficient denoising through image-specific training Mansour & Heckel (2023); Lequyer et al. (2022). Pixel2Pixel extends pair construction to non-local similar pixels, using a pixel bank and random sampling to generate multiple noisy views Ma et al. (2025). These pair-based approaches depend on content similarity and limited noise dependence between the constructed views. Deep Image Prior and Self2Self provide complementary approaches based on network structure and dropout Ulyanov et al. (2020); Quan et al. (2020). LoTA-N2N retains single-image training and uses a frozen teacher to estimate the interaction remaining in the sub-image objective.

Risk correction and relation to LoTA-N2N. Noise2Same bounds supervised risk using noisy reconstruction and a discrepancy between full and masked predictions Xie et al. (2020). UNSURE handles unknown noise levels through constraints on the denoiser's expected divergence Tachella et al. (2025). Learning to Recorrupt learns a noise transformation through a min–max objective that controls prediction–noise correlation Monroy et al. (2026). These works motivate explicit control of self-supervised objective mismatch. LoTA-N2N focuses on its spatial structure: it estimates residual–prediction interactions using a frozen single-image teacher and penalizes their regional absolute values. Our analysis explains why global averaging can hide large local interactions and quantifies how teacher error affects the resulting supervised-risk bound. Appendix A provides a detailed comparison of these methods and their assumptions.

3 Trace-Based View of Self-Supervision

Let $\mathbf{x}$ be a latent clean target, $\mathbf{u}$ a self-supervised input, and $\mathbf{s} = \mathbf{x}+\mathbf{r}$ a surrogate target with residual $\mathbf{r}$. We compare MSE reconstruction risks for the same input, with the input–target distribution held fixed as θ varies. Define $\mathbf{e}_\theta(\mathbf{u}) = f_\theta(\mathbf{u}) - \mathbf{x}$. Squared norms and Frobenius inner products are averages over scalar entries. The expected reconstruction risks satisfy

$$\mathcal{L}_{\mathrm{sup}}(\theta) = \mathbb{E}\|\mathbf{e}_\theta(\mathbf{u})\|_F^2, \tag{1}$$

$$\mathcal{L}_{\mathrm{ss}}(\theta) = \mathbb{E}\|f_\theta(\mathbf{u}) - \mathbf{s}\|_F^2 = \mathcal{L}_{\mathrm{sup}}(\theta) - 2\mathcal{T}(\theta) + C, \tag{2}$$

$$\mathcal{T}(\theta) = \mathbb{E}\langle \mathbf{r}, \mathbf{e}_\theta(\mathbf{u})\rangle_F, \qquad C = \mathbb{E}\|\mathbf{r}\|_F^2. \tag{3}$$

Thus self-supervised and supervised minimization agree whenever $\mathcal{T}(\theta)$ is zero for all θ. A sufficient condition is $\mathbb{E}[\mathbf{r} \mid \mathbf{u}, \mathbf{x}] = \mathbf{0}$, which covers the standard N2N argument. Blind spots seek conditional orthogonality by hiding target coordinates; re-corruption constructs independent noises under a compatible model; sub-image pairing combines approximate clean-neighbor similarity with noise independence. These constructions suppress the same interaction under different assumptions.

3.1 Revisiting Existing Methods Through the Trace Term

Equation (2) connects different MSE constructions through their surrogate residuals. The key distinction is whether the residual–prediction interaction is eliminated exactly or only approximately:

$$\begin{array}{lll} \text{N2N:} & \mathbf{r} = \mathbf{n}' & \mathbb{E}[\mathbf{r} \mid \mathbf{u}, \mathbf{x}] = \mathbf{0}, \\ \text{blind spot:} & \mathbf{r} = \mathbf{n}_J & \mathbb{E}\langle \mathbf{r}, \mathbf{e}_{\theta,J} \rangle_F = 0, \\ \text{NAC:} & \mathbf{r} = \mathbf{n}_o & |\mathcal{T}| \text{ is bounded by reused-noise energy}, \\ \text{R2R:} & \mathbf{r} = \mathbf{n} - \mathbf{B}\mathbf{z} & \mathrm{Cov}(\widehat{\mathbf{n}}, \mathbf{r} \mid \mathbf{x}) = \mathbf{0}, \\ \text{sub-images:} & \mathbf{r} = (\mathbf{x}_2 - \mathbf{x}_1) + \mathbf{n}_2 & \mathcal{T} \text{ is only approximately small.} \end{array} \tag{4}$$

This map compares MSE reconstruction terms; additional regularizers and inference procedures remain specific to each method. N2N and ideal blind spots achieve exact cancellation through conditional orthogonality, R2R recovers it under a compatible corruption model, and NAC and sub-image methods provide approximate control. LoTA-N2N estimates this remaining interaction and constrains its magnitude region by region.

Independent targets and blind spots. For N2N, $\mathbf{u} = \mathbf{x} + \mathbf{n}$, $\mathbf{s} = \mathbf{x} + \mathbf{n}'$, and $\mathbf{r} = \mathbf{n}'$. Conditional unbiasedness $\mathbb{E}[\mathbf{n}' \mid \mathbf{x}] = 0$ together with $\mathbf{n}' \perp\!\!\!\perp \mathbf{n} \mid \mathbf{x}$ implies $\mathbb{E}[\mathbf{n}' \mid \mathbf{u}, \mathbf{x}] = 0$. Neither Gaussianity nor equal input and target variances are required. For a blind-spot coordinate set J, the analogous interaction takes the form

$$\mathbb{E}\langle \mathbf{n}_J, f_{\theta,J}(\mathbf{y}_{J^c}) - \mathbf{x}_J \rangle_F . \tag{5}$$

It vanishes when the held-out noise is conditionally zero-mean and independent of context noise given the clean image. This explains both the benefit and limitations of hiding the center pixel: spatial correlation can violate the required orthogonality, while masking can discard useful context.

Noise2Same shares the goal of controlling self-supervised mismatch but uses a different surrogate. It bounds supervised risk through noisy reconstruction and a full-versus-masked prediction discrepancy, rather than estimating the residual–prediction interaction directly.

Weak-noise and model-based re-corruption. Noisy As Clean uses $\mathbf{u} = \mathbf{x} + \mathbf{n}_o + \mathbf{n}_s$ and target $\mathbf{s} = \mathbf{x} + \mathbf{n}_o$. Because the reused noise $\mathbf{n}_o$ also appears in the input, the conditional zero-mean argument generally fails. Instead, Cauchy–Schwarz gives the following bounds:

$$|\mathcal{T}(\theta)| \le \sqrt{\mathbb{E}\|\mathbf{n}_o\|_F^2} \sqrt{\mathbb{E}\|\mathbf{e}_\theta(\mathbf{u})\|_F^2}, \tag{6}$$

$$\|\nabla_\theta \mathcal{T}(\theta)\|_2 \le \sqrt{\mathbb{E}\|\mathbf{n}_o\|_F^2} \sqrt{\mathbb{E}\|\mathbf{J}_\theta f_\theta(\mathbf{u})\|_F^2}. \tag{7}$$

Weak reused noise therefore limits objective and gradient mismatch relative to prediction-error and Jacobian energy. R2R instead constructs $\widehat{\mathbf{y}} = \mathbf{y} + \mathbf{A}\mathbf{z}$ and $\widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{B}\mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is independent of the original noise and clean image. If the original noise is conditionally zero-mean Gaussian with covariance $\mathbf{\Sigma_x}$ and $\mathbf{A}\mathbf{B}^T = \mathbf{\Sigma_x}$, the constructed noises $\widehat{\mathbf{n}} = \mathbf{n} + \mathbf{A}\mathbf{z}$ and $\mathbf{r} = \mathbf{n} - \mathbf{B}\mathbf{z}$ have zero cross-covariance and are conditionally independent. Exact cancellation thus requires a compatible noise family and covariance model. Noisier2Noise and IDR similarly modify the input–target relation through additional or iterative corruption. Whenever a squared surrogate target is optimized, Eq. (2) identifies any residual interaction under the chosen corruption assumptions.

Sub-image residuals. For neighboring sub-images, choosing $\mathbf{x}_1$ as the clean target gives $\mathbf{r}_{1\to2} = \mathbf{y}_2 - \mathbf{x}_1 = (\mathbf{x}_2 - \mathbf{x}_1) + \mathbf{n}_2$. Pixelwise independent zero-mean noise can remove the noise interaction in expectation, while spatial redundancy is used to keep the clean-neighbor difference small. Textured edges and structured noise can violate these conditions. For Neighbor2Neighbor and ZSN2N, this view separates two sources of mismatch within pairwise MSE: clean-neighbor differences and residual-noise dependence. LoTA-N2N estimates their combined interaction with a frozen teacher and constrains it region by region, rather than assuming that either source disappears globally. The method does not eliminate assumptions; it replaces assumed cancellation with local estimates whose reliability depends on teacher estimation quality.

When differentiation and expectation can be interchanged, Eq. (2) gives the gradient relation

$$\nabla_\theta \mathcal{L}_{\mathrm{ss}} - \nabla_\theta \mathcal{L}_{\mathrm{sup}} = -2\nabla_\theta \mathcal{T}(\theta). \tag{8}$$

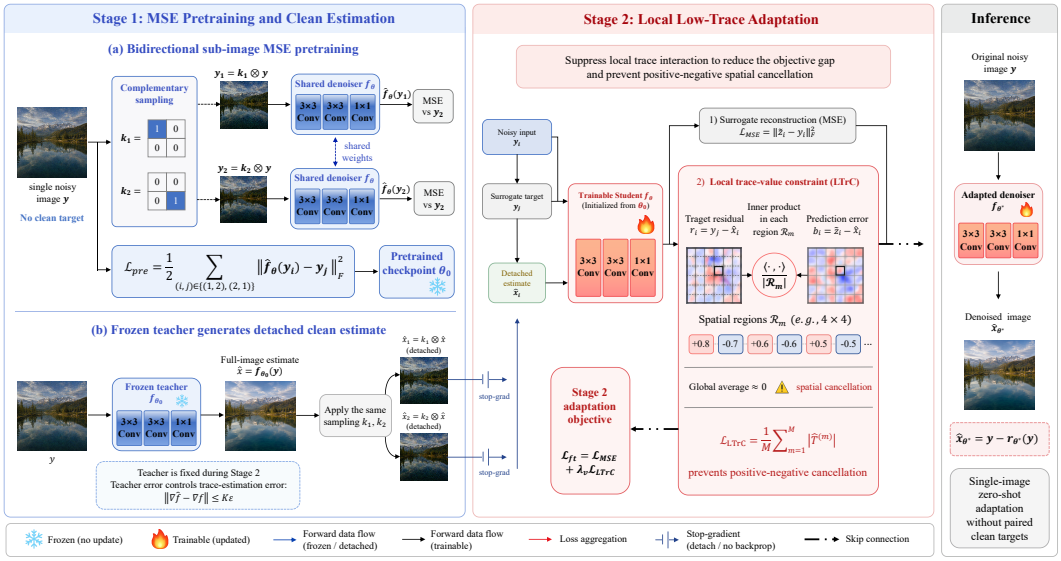


Figure 2: LoTA-N2N pipeline. Stage 1 trains a denoiser on complementary sub-images and freezes a copy as the teacher. Stage 2 adapts the student using bidirectional MSE and local trace correction, followed by full-image denoising with the adapted student. The schematic shows pairwise MSE and equal-sized regions; both stages also retain full-image consistency (Eq. (30)), and the local loss weights boundary regions by area.

A small value of $\mathcal{T}$ at a single checkpoint therefore does not imply a small gradient gap. We use Eq. (8) as a mechanism diagnostic rather than an optimization target: gradient-gap optimization requires double backpropagation and, as shown in Appendix D.6, provides no measurable gain in our reconstruction ablation.

4 LoTA-N2N

Given one noisy image $\mathbf{y}$, complementary 2×2 sampling kernels $\mathbf{k}_1$ and $\mathbf{k}_2$ form $\mathbf{y}_i = \mathbf{k}_i \otimes \mathbf{y}$, $i \in \{1, 2\}$. Because neighboring clean pixels may differ and their noise may be correlated, the resulting pair need not satisfy N2N assumptions.

Figure 2 summarizes the two-stage optimization. The frozen teacher supplies proxies during adaptation; the adapted student then denoises the full noisy image.

4.1 Stage 1: Detached clean proxies

Stage 1 minimizes bidirectional pair MSE (with the full-image consistency term detailed in Appendix C). Stage 2 initializes the student from the resulting checkpoint θ_0 and freezes a separate teacher copy to construct

$$\widehat{\mathbf{x}}_i = \mathrm{sg}\Big(\mathbf{k}_i \otimes \mathrm{clip}_{[0,1]}(f_{\theta_0}(\mathbf{y}))\Big), \qquad i \in \{1, 2\}, \tag{9}$$

where sg denotes stop-gradient and clipping keeps the teacher output in the image intensity range. For direction $i \to j$, define the clean-proxy error $\mathbf{e}_i = \widehat{\mathbf{x}}_i - \mathbf{x}_i$, prediction error $\mathbf{p}_{\theta,i} = f_\theta(\mathbf{y}_i) - \mathbf{x}_i$, and surrogate residual $\mathbf{q}_{i\to j} = \mathbf{y}_j - \mathbf{x}_i$. These definitions give the true interaction and its teacher-based estimate as follows:

$$\mathcal{T}_{i\to j} = \mathbb{E}\langle \mathbf{q}_{i\to j}, \mathbf{p}_{\theta,i}\rangle_F, \tag{10}$$

$$\widehat{\mathcal{T}}_{i\to j} = \mathbb{E}\langle \mathbf{y}_j - \widehat{\mathbf{x}}_i, f_\theta(\mathbf{y}_i) - \widehat{\mathbf{x}}_i\rangle_F. \tag{11}$$

Proxy reliability. If $\mathbb{E}\|\mathbf{e}_i\|_F^2 \le \epsilon^2$, $\mathbb{E}\|\mathbf{p}_{\theta,i}\|_F^2 \le P^2$, and $\mathbb{E}\|\mathbf{q}_{i\to j}\|_F^2 \le Q^2$, then expansion followed by Cauchy–Schwarz gives

$$|\widehat{\mathcal{T}}_{i\to j} - \mathcal{T}_{i\to j}| \le \epsilon(P + Q) + \epsilon^2. \tag{12}$$

The approximation is therefore useful only when Stage 1 supplies a sufficiently accurate proxy; the resulting bound makes this dependence quantitatively explicit.

4.2 Local trace correction

Partition each sub-image into M spatial regions Ω_m, including any smaller boundary regions. Let $w_m = |\Omega_m| / \sum_n |\Omega_n|$ denote their area weights. The estimated regional interaction is

$$\widehat{\mathcal{T}}_{i\to j}^{(m)} = \mathbb{E}\big[\langle (\mathbf{y}_j - \widehat{\mathbf{x}}_i)_{\Omega_m}, (f_\theta(\mathbf{y}_i) - \widehat{\mathbf{x}}_i)_{\Omega_m} \rangle_F\big], \tag{13}$$

where each inner product already averages over the scalar entries in its region. The bidirectional local trace correction is

$$\mathcal{L}_{\mathrm{LTrC}} = \frac{1}{2} \sum_{(i,j)\in\{(1,2),(2,1)\}} \sum_{m=1}^{M} w_m |\widehat{\mathcal{T}}_{i\to j}^{(m)}|. \tag{14}$$

For either direction, the triangle inequality yields

$$\left| \sum_m w_m \widehat{\mathcal{T}}_{i\to j}^{(m)} \right| \le \sum_m w_m |\widehat{\mathcal{T}}_{i\to j}^{(m)}|, \tag{15}$$

so regional absolute values prevent cancellation between regions and bound the absolute global estimate. Cancellation can still occur within individual regions.

Let $\mathcal{L}_{\mathrm{MSE}}$ denote bidirectional pair MSE plus the nonnegative full-image consistency loss retained in both training stages. Let $\mathcal{L}_{\mathrm{sup}}^{\mathrm{bi}}$ be the corresponding supervised sub-image risk and C_{bi} the average surrogate-residual energy. Define $\mathcal{L}_{\mathrm{LTrC}}^{\star}$ by replacing the estimated local interactions in Eq. (14) with their true values, and let $\Delta_{\mathcal{T}}$ be the average absolute estimation error. Applying Eq. (2) to pair MSE, adding the nonnegative consistency loss, and using Eq. (15) yields the following risk bound:

$$\mathcal{L}_{\mathrm{sup}}^{\mathrm{bi}} \le \mathcal{L}_{\mathrm{MSE}} - C_{\mathrm{bi}} + 2\mathcal{L}_{\mathrm{LTrC}}^{\star} \le \mathcal{L}_{\mathrm{MSE}} - C_{\mathrm{bi}} + 2\mathcal{L}_{\mathrm{LTrC}} + 2\Delta_{\mathcal{T}}. \tag{16}$$

Under the patch-wise conditions of Eq. (12), $\Delta_{\mathcal{T}} \le \epsilon(P+Q) + \epsilon^2$. We use the adaptation loss

$$\mathcal{L}_{\mathrm{ft}} = \mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{v}} \mathcal{L}_{\mathrm{LTrC}}, \tag{17}$$

where λ_{v} follows a cosine schedule. The proof and finite-sample implementation are given in Appendix C of the supplement.

Proof sketch and interpretation. For each direction, Eq. (2) gives $\mathcal{L}_{\mathrm{sup}}^{i} = \mathcal{L}_{\mathrm{ss}}^{i\to j} - C_{i\to j} + 2\sum_m w_m \mathcal{T}_{i\to j}^{(m)}$. Taking regional absolute values, averaging directions, and adding the consistency loss proves the first inequality in Eq. (16). The second follows from $|\mathcal{T}^{(m)}| \le |\widehat{\mathcal{T}}^{(m)}| + |\widehat{\mathcal{T}}^{(m)} - \mathcal{T}^{(m)}|$. Finally, expanding the estimated interaction around $\widehat{\mathbf{x}}_i = \mathbf{x}_i + \mathbf{e}_i$ produces three terms bounded respectively by ϵP, ϵQ, and ϵ^2. The coefficient two therefore has a population-risk interpretation, whereas the empirical weight absorbs normalization and finite-sample effects. The result bounds expected sub-image risk; it does not guarantee gradient alignment or a corresponding bound on full-image inference risk.

5 Experiments

5.1 Setup

We evaluate Kodak24, McMaster18 Zhang et al. (2011), Set14 Zeyde et al. (2010), and Set5 Bevilacqua et al. (2012), plus confocal images from FMD Zhang et al. (2019) and pediatric chest X-rays Kermany et al. (2018). Synthetic tests cover Gaussian, Poisson, and mixed Poisson–Gaussian noise. Each noisy image is adapted independently; clean images are used only for evaluation. All rerun methods receive identical saved noisy inputs, clipping, and PSNR/SSIM code Wang et al. (2004). The default backbone is a 0.022M-parameter, three-layer residual CNN. Stage 1 runs for 3,000 Adam steps. Stage 2 starts from the same checkpoint for every matched comparison and runs for 800 steps with region width 32, $\lambda_{\mathrm{v}} = 8$, and a cosine decay to 15% of the initial trace weight. Per-image statistics, additional baselines, noise definitions, and implementation details are in Appendix D.

Method	Mode	Kodak24			McMaster18			Set14			Set5		
		Noise level			Noise level			Noise level			Noise level		
Gaussian noise		$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$
DnCNN	Offline	**35.33**	**33.17**	**31.50**	**34.48**	**32.89**	**31.51**	**33.65**	**31.99**	**30.68**	**35.01**	**33.26**	**31.83**
N2N	Offline	34.86	32.88	31.46	33.57	32.22	31.06	33.07	31.63	30.49	34.61	33.07	31.81
CBM3D	Online	33.50	31.30	29.83	34.49	32.18	30.48	32.92	30.74	29.22	**34.38**	**32.85**	**31.67**
DIP2000	Online	33.13	31.13	29.69	33.65	31.86	30.34	29.91	29.26	28.26	30.50	29.67	28.83
Noise2Self	Online	28.80	28.23	27.44	30.46	29.64	28.62	29.58	28.37	28.21	30.52	29.23	27.95
Noise2Fast	Online	32.22	30.78	29.63	33.89	32.10	30.64	31.49	30.08	28.93	33.11	31.58	30.33
ZSN2N	Online	33.91	31.98	30.43	34.19	32.00	30.31	33.00	31.02	29.55	33.82	31.88	30.32
LoTA-N2N	Online	**34.46**	**32.39**	**30.87**	**34.58**	**32.41**	**30.77**	**33.12**	**31.23**	**29.84**	34.03	32.16	30.67
Poisson noise		$\lambda = 60$	$\lambda = 50$	$\lambda = 40$	$\lambda = 60$	$\lambda = 50$	$\lambda = 40$	$\lambda = 60$	$\lambda = 50$	$\lambda = 40$	$\lambda = 60$	$\lambda = 50$	$\lambda = 40$
DnCNN	Offline	**31.80**	**31.35**	**30.68**	24.03	23.89	23.67	26.88	26.61	26.20	23.39	23.26	23.06
N2N	Offline	28.25	28.83	29.37	**28.56**	**29.18**	**29.80**	**27.09**	**27.81**	**28.33**	**27.98**	**28.38**	**29.37**
CBM3D	Online	28.33	28.26	28.08	29.33	29.21	28.97	29.37	28.79	28.21	28.74	28.10	27.38
DIP2000	Online	29.11	28.62	28.04	30.29	29.78	29.33	26.12	25.83	25.48	28.70	28.26	27.65
Noise2Self	Online	27.08	26.77	26.67	29.03	29.00	28.31	27.75	27.24	27.00	27.93	26.94	26.56
Noise2Fast	Online	29.29	28.87	28.37	31.01	30.54	29.98	28.61	28.20	27.71	30.41	29.98	29.42
ZSN2N	Online	30.36	**29.93**	29.28	30.80	30.47	29.86	29.09	28.63	27.98	30.36	29.82	29.25
LoTA-N2N	Online	**30.45**	**29.93**	**29.29**	**31.35**	**30.84**	**30.22**	**29.43**	**28.92**	**28.31**	**30.67**	**30.17**	**29.54**
Mixed noise		$p = 60$ $r = 2$	$p = 50$ $r = 2$	$p = 40$ $r = 2$	$p = 60$ $r = 2$	$p = 50$ $r = 2$	$p = 40$ $r = 2$	$p = 60$ $r = 2$	$p = 50$ $r = 2$	$p = 40$ $r = 2$	$p = 60$ $r = 2$	$p = 50$ $r = 2$	$p = 40$ $r = 2$
DnCNN	Offline	**31.72**	**31.24**	**30.53**	21.80	21.72	21.57	25.94	25.68	25.31	20.76	20.68	20.53
N2N	Offline	29.08	29.53	29.59	**29.46**	**29.98**	**30.28**	**28.10**	**28.60**	**28.74**	**28.98**	**29.42**	**30.06**
CBM3D	Online	30.37	29.85	29.22	29.00	28.35	27.59	29.24	28.71	28.19	28.80	28.12	27.32
DIP2000	Online	26.63	26.35	25.97	28.15	27.97	27.56	26.11	25.81	25.49	28.55	28.19	27.59
Noise2Self	Online	28.24	28.17	27.79	30.10	29.86	29.63	27.89	27.40	26.92	27.95	27.51	26.26
Noise2Fast	Online	29.38	28.97	28.45	31.14	30.70	**30.14**	28.58	28.17	27.67	30.40	29.99	29.41
ZSN2N	Online	30.03	29.50	28.85	30.73	30.23	29.58	29.08	28.57	27.95	30.33	29.84	29.24
LoTA-N2N	Online	**30.44**	**29.92**	**29.27**	**31.27**	**30.76**	**30.14**	**29.38**	**28.89**	**28.28**	**30.62**	**30.17**	**29.52**

Table 1: Quantitative comparison of different denoising methods on Kodak24, McMaster18, Set14, and Set5. The evaluation metric is PSNR (dB). The best offline result and the best online result in each setting are highlighted separately in bold; ties at the reported precision are all highlighted. In the mixed-noise block, p denotes the Poisson peak value and r denotes the Gaussian read-noise standard deviation on the 0–255 intensity scale. DnCNN and N2N are randomly initialized dataset-trained methods: models trained on Kodak24 are evaluated on McMaster18, Set14, and Set5, while models trained on McMaster18 are evaluated on Kodak24. No official pretrained weights are used. All newly reproduced entries use the same saved noisy inputs as LoTA-N2N.

5.2 Main comparison

Table 1 reports the complete natural-image comparison rather than a selected subset. Offline dataset-trained references and online per-image methods are separated because they use different supervision regimes. Among online methods, LoTA-N2N is best in nine of the twelve Gaussian settings, best or tied in all twelve Poisson settings, and best or tied throughout the mixed-noise block. Figure 3 provides two qualitative, multi-method comparisons under mixed noise.

The controlled Confocal and X-ray results are reported in Appendix Table 6. LoTA-N2N achieves the highest PSNR in all eight listed settings under the evaluated Gaussian and Poisson corruptions.

The improvement is not an artifact of an extra 800 updates. Against ZSN2N-3800, which receives the identical Stage 2 budget using MSE alone, mean gains over three levels are $+0.37$, $+0.39$, and $+0.39$ dB on Kodak24 and $+0.50$, $+0.58$, and $+0.58$ dB on McMaster18 for Gaussian, Poisson, and mixed noise. All six image-level bootstrap intervals exclude zero, with win rates from 91.7% to 100%. On Set14, Poisson and mixed gains also remain significant. The complete confidence intervals and win rates are reported in Appendix Table 7. Appendix Figure 4 complements the table with representative Kodak24 distributions, each showing gains on 23 of 24 images.

5.3 Noise shifts and mechanism tests

We next test IID Gaussian noise ($\sigma = 25/255$), spatially varying Gaussian noise (standard deviation $5/255$ to $50/255$), and mixed Poisson–Gaussian noise (peak 30 and read noise $5/255$) on Kodak24.

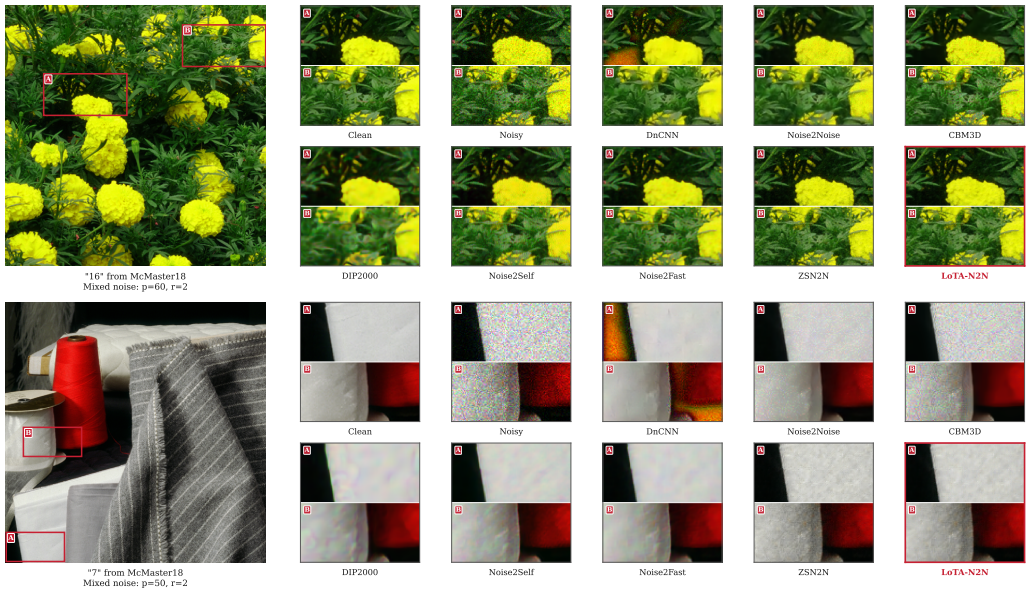


Figure 3: Qualitative comparison on two McMaster18 images under mixed Poisson–Gaussian noise. Each example shows the full image, two enlarged regions, the clean and noisy references, and outputs from DnCNN, Noise2Noise, CBM3D, DIP2000, Noise2Self, Noise2Fast, ZSN2N, and LoTA-N2N. Table 1 provides the quantitative comparisons.

Stage 2 objective	PSNR/SSIM	Global trace estimate	$\mathcal{L}_{\mathrm{LTrC}}$	$G_{\mathrm{rel}} \downarrow$	$G_{\cos} \uparrow$
No Stage 2	34.14 / 0.8900	–	–	–	–
MSE, iteration-matched	34.17 / 0.8912	4.11×10^{-5}	5.14×10^{-5}	1.008	0.058
MSE + Global TrC	34.49 / 0.8985	5.53×10^{-6}	3.26×10^{-5}	0.961	0.251
MSE + LTrC	34.58 / 0.9027	1.37×10^{-5}	2.02×10^{-5}	0.960	0.164
MSE + GTrC	34.17 / 0.8912	4.11×10^{-5}	5.14×10^{-5}	1.009	0.053
MSE + LTrC + GTrC	34.58 / 0.9026	1.37×10^{-5}	2.02×10^{-5}	0.959	0.171
Oracle trace diagnostic†	– / –	$\lvert\overline{\mathcal{T}}\rvert = 1.00\times10^{-4}$	$E_{\mathcal{T}} = 1.169$	–	–

Table 2: Objective ablation and mechanism diagnostics on McMaster18 with Gaussian noise $\sigma = 10$. Entries are means over all 18 images, and all Stage 2 trace regularizers use a floor of 0.15. The MSE control uses exactly the same Stage 2 iterations as the trace-regularized variants. MSE includes bidirectional pairing and full-image consistency; GTrC rows are diagnostic ablations. The global estimate averages absolute directional traces. †The oracle row instead averages the absolute signed bidirectional trace and the per-image relative estimation error, defined in Appendix D.6, at the final LTrC checkpoint. The mechanism study uses separately trained Stage 1 checkpoints.

Appendix Table 8 shows that the gain over both Noise2Fast and ZSN2N persists when no method receives the noise map or corruption parameters. Relative to ZSN2N, LoTA-N2N gains 0.42, 0.31, and 0.41 dB; all corresponding differences are computed directly from the reported means. LoTA-N2N achieves the highest PSNR in all three conditions and the highest SSIM under IID and spatially varying Gaussian noise among the compared methods.

Table 2 isolates the Stage 2 objective on McMaster18 at $\sigma = 10$. LTrC improves the iteration-matched MSE control from 34.17 to 34.58 dB while reducing the estimated mean local interaction from 5.14×10^{-5} to 2.02×10^{-5}. Global TrC reaches 34.49 dB, whereas GTrC is indistinguishable from MSE and adds no benefit to LTrC at the reported precision. The gradient quantities are therefore diagnostics rather than training terms.

Setting	Bidirectional + consistency	Residual	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 25$
A_1	✘	✘	33.7612 / 0.8826	31.5647 / 0.8262	29.9177 / 0.7735	28.5585 / 0.7199
A_2	✔	✘	34.0974 / 0.8862	32.0574 / 0.8429	30.5368 / 0.8039	29.2653 / 0.7660
A_3	✘	✔	34.1347 / 0.8944	31.8502 / 0.8380	30.1274 / 0.7817	28.7153 / 0.7282
$\mathbf{A}_4$	✔	✔	**34.5810 / 0.9026**	**32.4209 / 0.8585**	**30.7546 / 0.8156**	**29.3918 / 0.7748**

Table 3: Design ablation on McMaster18 under the final LTrC-only method. Entries report mean PSNR/SSIM over all 18 images using one fixed noise realization per image and optimization seed 2027. Every setting trains its own Stage 1 model and receives an 800-step floor-0.15 Stage 2 update with $\mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{v}}\mathcal{L}_{\mathrm{LTrC}}$, using the pair and consistency terms enabled in that setting. Design-ablation checkpoints are separate from the main benchmark despite matching noise and model seeds, explaining the small deviations from the main benchmark.

Backbone	Params (M)	Gaussian		Poisson		Mixed Noise		Correlated Gaussian		Time overhead
		MSE	LoTA-N2N	MSE	LoTA-N2N	MSE	LoTA-N2N	MSE	LoTA-N2N	vs. standard
Three-layer CNN (standard)	0.022	30.19	**30.77**	30.28	**30.84**	28.11	**28.71**	**23.08**	22.98	1.00×
Eight-layer residual CNN	0.126	30.52	**30.60**	30.32	**30.52**	28.52	**28.57**	23.36	**23.38**	2.88× (+188.4%)
Compact encoder–decoder	0.053	30.85	**31.00**	30.98	**31.08**	28.97	**29.00**	**23.12**	23.00	1.57× (+57.5%)

Table 4: Transfer of the proposed objective across denoising backbones on McMaster18. Entries report PSNR (dB) for Gaussian noise with $\sigma = 20$, Poisson noise with peak value 50, mixed Poisson–Gaussian noise with peak value 50 plus Gaussian noise $\sigma = 20$, and correlated Gaussian noise with $\sigma = 20$ using a 3×3 kernel with kernel standard deviation 0.8. All LoTA-N2N entries use the final floor-0.15 configuration. Runtime is normalized by the standard three-layer CNN configuration.

Table 3 evaluates bidirectional-consistency training and residual parameterization. Their combination gives the highest PSNR/SSIM at every reported Gaussian level. Bidirectionality and consistency are switched together, so their individual effects are not isolated by this comparison.

5.4 Architecture and cost

Table 4 repeats the matched objective comparison with three backbones. LTrC improves every backbone under IID Gaussian, Poisson, and mixed noise. The residual CNN gains 0.05–0.20 dB across these settings, while the compact encoder–decoder gains 0.03–0.15 dB. These improvements show that local interaction control transfers beyond the standard lightweight CNN.

At 500×500, LoTA-N2N takes 20.37 seconds and 0.28 GB on H100, versus 20.24 seconds and 0.28 GB for ZSN2N at roughly 3,800 updates (Appendix D.7).

5.5 Scope and limitations

LoTA-N2N avoids specifying a noise distribution, while relying on teacher quality and sub-image redundancy as quantified by Eq. (12). The trace-weight floor was benchmark-tuned; Appendix D.7 documents the selection process and subsequent Set12 sensitivity study.

6 Conclusion

We revisited MSE self-supervision by linking existing assumptions to a common trace term. Eliminating or limiting this term brings surrogate-target training closer to its supervised counterpart, motivating explicit control when the interaction persists. LoTA-N2N implements this principle using frozen-teacher clean proxies and regional absolute trace penalties, which prevent cancellation between regions. A supervised sub-image risk bound accounts for teacher-estimation error and motivates the practical adaptation loss. Experiments show gains over MSE controls under matched optimization budgets. Among the reported online methods, LoTA-N2N is best or tied in 21 of 24 Gaussian and Poisson settings and all 12 mixed-noise settings. These results support local interaction control as an effective objective for single-image denoising under the evaluated noise conditions.

Reproducibility statement. Appendix D specifies datasets, corruption protocols, optimization schedules, and matched evaluation; Appendix C provides the complete derivations; and Appendix D.7 reports the selected configuration and resolution-scaling runtime results. Code and fixed noisy inputs should be released with the anonymous supplementary submission package for reviewer verification during peer review.

AI USE STATEMENT

Generative AI was used for language editing, structural reorganization, and LaTeX formatting of this manuscript. It was not used to generate experimental measurements or to replace the authors' verification of derivations, citations, or reported results. The authors reviewed all AI-assisted work and take responsibility for the final content of this paper.

REFERENCES

Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3155–3164, 2019.

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533, 2019.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pp. 135.1–135.10, 2012.

Hamadi Chihaoui and Paolo Favaro. Masked and shuffled blind spot denoising for real-world images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3025–3034, 2024.

Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1712–1722, 2019.

Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14781–14790, 2021.

Bo Jiang, Yao Lu, Jiahuan Wang, Guangming Lu, and David Zhang. Deep image denoising with adaptive priors. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5124–5136, 2022. doi: 10.1109/TCSVT.2022.3149518.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. doi: 10.1016/j.cell.2018.02.010.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void: Learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2129–2137, 2019.

Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. AP-BSN: Self-supervised denoising for real-world images via asymmetric PD and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17725–17734, 2022.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2965–2974, 2018.

Jason Lequyer, Reuben Philip, Amit Sharma, and Laurence Pelletier. Noise2Fast: Fast self-supervised single image blind denoising. *Nature Machine Intelligence*, 4:953–963, 2022. doi: 10.1038/s42256-022-00547-8.

Qing Ma, Junjun Jiang, Xiong Zhou, Pengwei Liang, Xianming Liu, and Jiayi Ma. Pixel2Pixel: A pixelwise approach for zero-shot single image denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4614–4629, 2025. doi: 10.1109/TPAMI.2025.3546870.

Youssef Mansour and Reinhard Heckel. Zero-shot Noise2Noise: Efficient image denoising without any data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14018–14027, 2023. doi: 10.1109/CVPR52729.2023.01347.

Brayan Monroy, Jorge Bacca, and Julián Tachella. Generalized recorrupted-to-recorrupted: Self-supervised learning beyond gaussian noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28155–28164, 2025.

Brayan Monroy, Jorge Bacca, and Julián Tachella. Learning to recorrupt: Noise distribution agnostic self-supervised image denoising. *arXiv preprint arXiv:2603.25869*, 2026. URL https://arxiv.org/abs/2603.25869.

Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2Noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12064–12072, 2020.

Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2043–2052, 2021. doi: 10.1109/CVPR46437.2021.00208.

Junyoung Park, Youngjin Oh, and Nam Ik Cho. TM-BSN: Triangular-masked blind-spot network for real-world self-supervised image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29877–29886, 2026.

Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2self with dropout: Learning self-supervised denoising from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1890–1898, 2020.

Yuhui Quan, Tianxiang Zheng, Zhiyuan Ma, and Hui Ji. Zero-shot blind-spot image denoising via implicit neural sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7502–7512, 2025.

Julián Tachella, Mike Davies, and Laurent Jacques. UNSURE: Self-supervised learning with unknown noise level and stein's unbiased risk estimate. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ScVnYBaSEw.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128:1867–1888, 2020. doi: 10.1007/s11263-020-01303-4.

Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2027–2036, 2022. doi: 10.1109/CVPR52688.2022.00207.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Yaochen Xie, Zhengyang Wang, and Shuiwang Ji. Noise2Same: Optimizing a self-supervised bound for image denoising. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jun Xu, Yuan Huang, Ming-Ming Cheng, Li Liu, Fan Zhu, Zhou Xu, and Ling Shao. Noisy-as-clean: Learning self-supervised denoising from corrupted image. *IEEE Transactions on Image Processing*, 29:9316–9329, 2020. doi: 10.1109/TIP.2020.3026622.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, volume 6920 of *Lecture Notes in Computer Science*, pp. 711–730, 2010.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. doi: 10.1109/TIP.2017.2662206.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. doi: 10.1109/TIP.2018.2839891.

Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. The McMaster colour image dataset. McMaster University and Hong Kong Polytechnic University, 2011. URL https://www4.comp.polyu.edu.hk/~cslzhang/CDM_Dataset.htm.

Yi Zhang, Dasong Li, Ka Lung Law, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. IDR: Self-supervised image denoising via iterative data refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2098–2107, 2022.

Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11710–11718, 2019.

SUPPLEMENTARY MATERIAL

This supplement provides the complete derivations, experimental protocols, cross-domain evaluations, paired statistics, and mechanism diagnostics supporting the main paper.

A EXTENDED RELATED-WORK ANALYSIS

A.1 SUPERVISED AND REPEATED-OBSERVATION DENOISING

Supervised denoisers learn from aligned noisy–clean image pairs and have achieved strong performance with increasingly expressive network architectures Zhang et al. (2017; 2018); Jiang et al. (2022); Anwar & Barnes (2019); Guo et al. (2019). Their applicability is limited when clean references are unavailable or hard to align with noisy observations.

Noise2Noise (N2N) Lehtinen et al. (2018) replaces each clean target with an independently corrupted observation of the same latent image. Under conditionally independent and unbiased target noise, its expected MSE differs from supervised MSE only by a parameter-independent constant. This removes the need for clean targets but still requires aligned repeated acquisitions, often unavailable in microscopy and single-shot imaging.

A.2 SINGLE-IMAGE SELF-SUPERVISED DENOISING

Blind-spot methods construct supervision by predicting selected pixels from their spatial context. Noise2Void Krull et al. (2019) and Noise2Self Batson & Royer (2019) implement this principle through masking, while later methods use restricted receptive fields or asymmetric architectures Laine et al. (2019); Lee et al. (2022); Wang et al. (2022). Their theoretical justification commonly relies on conditionally zero-mean noise and independence between the held-out noise and the observed context. Spatial correlation can violate this condition, while excluding the center pixel may also remove useful information.

Noise2Same Xie et al. (2020) relaxes strict $\mathcal{J}$-invariance by deriving a self-supervised upper bound on supervised denoising risk. Its objective combines full-image noisy reconstruction with a prediction-invariance discrepancy between full and masked inputs. Noise2Same therefore does not require the denoiser itself to be $\mathcal{J}$-invariant or require the noise distribution to be supplied to the algorithm, although the theoretical bound assumes zero-mean IID noise across image coordinates. It shares LoTA-N2N's upper-bound motivation, but controls a masking-based discrepancy rather than directly estimating residual–prediction interactions.

Sub-image methods construct proxy pairs from neighboring pixels within one noisy image. Neighbor2Neighbor Huang et al. (2021) uses neighboring subsampling, while Zero-Shot Noise2Noise Mansour & Heckel (2023) and Noise2Fast Lequyer et al. (2022) enable efficient per-image optimization without an external training set. These approaches are well suited to zero-shot denoising, but their proxy pairs only approximately satisfy N2N assumptions: clean neighbors may differ, and their noise can remain statistically dependent.

Re-corruption methods instead synthesize additional noisy observations. Noisy As Clean Xu et al. (2020), Noisier2Noise Moran et al. (2020), Recorrupted-to-Recorrupted Pang et al. (2021), and IDR Zhang et al. (2022) construct noisier inputs or iterative targets using assumptions about the corruption process. Deep Image Prior Ulyanov et al. (2020) and Self2Self Quan et al. (2020) instead exploit network bias, dropout, and image-specific optimization without requiring paired training images.

A.3 RELATION TO LOTA-N2N

Table 5 revisits representative methods through the discrepancy between self-supervised and supervised MSE. For the noisy-target constructions analyzed in Appendix B, the assumptions eliminate or limit the residual–prediction interaction, making the two objectives equal or closer up to a parameter-independent constant. This interpretation motivates estimating and penalizing the remaining interaction explicitly. LoTA-N2N uses frozen-teacher proxies for this purpose and applies the penalty locally to prevent cancellation between regions.

Method family	Observable construction	Mechanism for controlling objective mismatch	Principal condition or limitation
N2N	Independent noisy target	Makes the target residual conditionally mean-zero	Requires registered repeated observations with conditionally independent and unbiased target noise
Blind spot	Held-out noisy pixels	Excludes target coordinates so that suitable noise assumptions make the residual orthogonal to prediction error	Relies on conditionally mean-zero, context-independent noise; spatial correlation breaks the argument
Noise2Same	Full and masked predictions	Upper-bounds supervised risk using noisy reconstruction and a prediction-invariance discrepancy	Its bound assumes zero-mean IID coordinate noise; it uses a masking-based surrogate rather than direct interaction estimation
NAC	Observed noise reused in the target	Bounds the interaction through reused-noise and prediction-error energies	Provides no exact zero-interaction guarantee when the observed corruption is strong
R2R/IDR	Model-based or iteratively refined corrupted targets	Uses re-corruption assumptions or target refinement to reduce the resulting interaction	Requires a compatible corruption model or sufficiently reliable iterative targets
Sub-image pairs	Neighboring-pixel proxy targets	Combines approximate clean-neighbor similarity with noise independence	Texture, clean-neighbor mismatch, and correlated noise can leave a nonzero interaction
LoTA-N2N	Frozen-teacher clean proxies	Estimates the residual–prediction interaction and penalizes its local absolute magnitude	Depends on teacher quality and sub-image redundancy rather than a prescribed corruption model

Table 5: Representative self-supervised denoising methods viewed through objective mismatch control. N2N, blind-spot, weak-noise, re-corruption, and sub-image methods suppress the residual–prediction interaction exactly or approximately through different assumptions. Noise2Same controls a related supervised-risk bound through prediction invariance; LoTA-N2N estimates and locally suppresses the interaction during adaptation.

Noise2Same already established a supervised-risk upper bound without strict $\mathcal{J}$-invariance. Its full-versus-masked prediction discrepancy differs from the trace estimate used here. LoTA-N2N instead performs per-image adaptation on complementary sub-images, using a frozen teacher to estimate the interaction and regional absolute penalties to prevent cancellation.

This comparison concerns the MSE reconstruction term, not an equivalence between complete algorithms. The methods retain distinct data requirements, architectures, and auxiliary losses; the trace analysis explains how their surrogate targets relate to supervised reconstruction.

LoTA-N2N does not require a predefined noise distribution or re-corruption model. Its reliability instead depends on Stage 1 proxy quality and the spatial redundancy of complementary sub-images. Appendix C bounds proxy-estimation error and derives the supervised-risk guarantee.

B Complete Trace-Based Analysis

B.1 Objective and Gradient Discrepancies

Let $\mathbf{x}$ be a latent clean target, $\mathbf{u}$ a self-supervised input, and $\mathbf{s} = \mathbf{x}+\mathbf{r}$ a surrogate target with residual $\mathbf{r}$. We compare reconstruction risks for the same input and hold the input–target distribution fixed as θ varies. This analysis applies to the MSE reconstruction term, independently of any additional regularizers used by a method. Define $\mathbf{e}_\theta(\mathbf{u}) = f_\theta(\mathbf{u}) - \mathbf{x}$. Squared norms and Frobenius inner products average over scalar entries. The expected reconstruction losses take the following form:

$$\begin{aligned} \mathcal{L}_{\mathrm{sup}}(\theta) &= \mathbb{E}\|\mathbf{e}_\theta(\mathbf{u})\|_F^2, \\ \mathcal{L}_{\mathrm{ss}}(\theta) &= \mathbb{E}\|f_\theta(\mathbf{u}) - \mathbf{s}\|_F^2. \end{aligned} \tag{18}$$

For matrices with d scalar entries, the normalized inner product is $\langle \mathbf{A}, \mathbf{B}\rangle_F = d^{-1}\,\mathrm{Tr}(\mathbf{A}^T\mathbf{B})$. Expanding the squared difference motivates the trace term

$$\mathcal{T}(\theta) = \mathbb{E}[\langle \mathbf{r}, \mathbf{e}_\theta(\mathbf{u})\rangle_F]\,. \tag{19}$$

The reconstruction losses therefore satisfy

$$\mathcal{L}_{\mathrm{ss}}(\theta) = \mathcal{L}_{\mathrm{sup}}(\theta) - 2\mathcal{T}(\theta) + C, \tag{20}$$

where $C = \mathbb{E}\|\mathbf{r}\|_F^2$ is independent of θ. Thus, after removing this constant, the objective discrepancy has magnitude $2|\mathcal{T}(\theta)|$. Assumptions that eliminate or bound the trace therefore recover exact or approximate agreement with supervised MSE.

Proposition 1 (Gradient discrepancy). If f_θ is differentiable with respect to θ and differentiation can be interchanged with expectation, then

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\mathrm{ss}} - \nabla_\theta \mathcal{L}_{\mathrm{sup}} &= -2\nabla_\theta \mathcal{T}(\theta), \\ \|\nabla_\theta \mathcal{L}_{\mathrm{ss}} - \nabla_\theta \mathcal{L}_{\mathrm{sup}}\|_2 &= 2\|\nabla_\theta \mathcal{T}(\theta)\|_2. \end{aligned} \tag{21}$$

Proof. Differentiate the exact decomposition. The constant C has zero derivative because the target construction is independent of the model parameters.

Proposition 1 separates objective-value agreement from optimization-direction agreement. A small trace value at a single parameter point does not by itself imply a small gradient discrepancy. In N2N, conditionally independent and unbiased target noise instead yields $\mathcal{T}(\theta) \equiv 0$ for all θ, and hence $\nabla_\theta \mathcal{T}(\theta) \equiv 0$. We therefore use gradient quantities to diagnose the mechanism. The trainable objective developed below combines reconstruction MSE with local trace magnitudes to bound expected supervised sub-image risk, up to estimation error, using only first-order derivatives during training.

B.2 Revisiting Existing Methods Through the Trace Term

The preceding decomposition provides a common criterion for MSE-based self-supervision. In particular, a convenient sufficient condition for exact risk equivalence up to a constant is

$$\mathbb{E}[\mathbf{r} \mid \mathbf{u}, \mathbf{x}] = \mathbf{0}. \tag{22}$$

Indeed, conditioning on $(\mathbf{u}, \mathbf{x})$ gives

$$\mathcal{T}(\theta) = \mathbb{E}_{\mathbf{u},\mathbf{x}} \left\langle \mathbb{E}[\mathbf{r} \mid \mathbf{u}, \mathbf{x}], f_\theta(\mathbf{u}) - \mathbf{x} \right\rangle_F = 0 \tag{23}$$

for every θ. Consequently, $\nabla_\theta \mathcal{T}(\theta) = 0$ and the self-supervised and supervised gradients coincide. Equation (22) is sufficient rather than necessary; a weaker sufficient condition is $\mathbb{E}\langle \mathbf{r}, \mathbf{e}_\theta(\mathbf{u}) \rangle_F = 0$ for all θ. This distinction reveals that several apparently different methods suppress the same interaction through different data constructions or statistical assumptions. Table 5 summarizes the common mechanism and the distinct sufficient conditions.

Independent noisy targets. For N2N, $\mathbf{u} = \mathbf{x} + \mathbf{n}$, $\mathbf{s} = \mathbf{x} + \mathbf{n}'$, and $\mathbf{r} = \mathbf{n}'$. If $\mathbb{E}[\mathbf{n}' \mid \mathbf{x}] = 0$ and $\mathbf{n}' \perp\!\!\!\perp \mathbf{n} \mid \mathbf{x}$, then $\mathbb{E}[\mathbf{n}' \mid \mathbf{u}, \mathbf{x}] = 0$, so $\mathcal{T}(\theta) \equiv 0$. Gaussianity and equal noise variances are not essential to this argument; conditional unbiasedness and independence are the key requirements. The practical restriction is the need for aligned repeated observations whose target corruption is independent of the input corruption.

Blind-spot prediction. For a held-out coordinate set J, blind-spot methods predict $\mathbf{y}_J = \mathbf{x}_J + \mathbf{n}_J$ using only $\mathbf{y}_{J^c}$. Their trace interaction is

$$\mathbb{E} \left\langle \mathbf{n}_J, f_{\theta,J}(\mathbf{y}_{J^c}) - \mathbf{x}_J \right\rangle_F. \tag{24}$$

The $\mathcal{J}$-invariant construction makes this term zero when the held-out noise is conditionally mean-zero and independent of the observed-coordinate noise given the clean image Batson & Royer (2019). Thus, masking is not merely an architectural device: in the unified view, it is a mechanism for enforcing trace orthogonality, which spatial noise correlation can invalidate.

Weak-noise targets. NAC forms $\mathbf{u} = \mathbf{x} + \mathbf{n}_o + \mathbf{n}_s$ and uses $\mathbf{s} = \mathbf{x} + \mathbf{n}_o$ as the target. Here the target residual $\mathbf{n}_o$ is also contained in the input, so Eq. (22) generally does not hold. Applying Cauchy–Schwarz yields the following bounds:

$$\begin{aligned} |\mathcal{T}(\theta)| &\le \sqrt{\mathbb{E}\|\mathbf{n}_o\|_F^2} \sqrt{\mathbb{E}\|\mathbf{e}_\theta(\mathbf{u})\|_F^2}, \\ \|\nabla_\theta \mathcal{T}(\theta)\|_2 &\le \sqrt{\mathbb{E}\|\mathbf{n}_o\|_F^2} \sqrt{\mathbb{E}\|\mathbf{J}_\theta f_\theta(\mathbf{u})\|_F^2}, \end{aligned} \tag{25}$$

where $\mathbf{J}_\theta f_\theta(\mathbf{u})$ is the Jacobian of the network output with respect to its parameters. The weak-noise assumption limits these discrepancies through reused-noise energy, relative to the prediction-error and Jacobian energies in the bounds.

Model-based re-corruption. R2R starts from $\mathbf{y} = \mathbf{x} + \mathbf{n}$ and constructs

$$\widehat{\mathbf{y}} = \mathbf{y} + \mathbf{A}\mathbf{z}, \qquad \widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{B}\mathbf{z}. \tag{26}$$

For $\mathbf{n} \mid \mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma_x})$, $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ independent of $(\mathbf{x}, \mathbf{n})$, and $\mathbf{A}\mathbf{B}^T = \mathbf{\Sigma_x}$, the constructed noises satisfy $\mathrm{Cov}(\widehat{\mathbf{n}}, \widetilde{\mathbf{n}} \mid \mathbf{x}) = \mathbf{\Sigma_x} - \mathbf{A}\mathbf{B}^T = 0$. Joint Gaussianity then turns zero covariance into conditional independence, again making the trace term vanish. The cost of this exact cancellation is a specified noise family and a compatible covariance model.

Sub-image pairs and the role of LoTA-N2N. For neighboring sub-images, taking $\mathbf{x}_1$ as the clean target gives $\mathbf{r}_{1\to2} = \mathbf{y}_2 - \mathbf{x}_1 = (\mathbf{x}_2 - \mathbf{x}_1) + \mathbf{n}_2$. Pixelwise independent, zero-mean noise can remove the noise contribution in expectation, while local similarity is used to keep $\mathbf{x}_2 - \mathbf{x}_1$ small. Neither property is guaranteed for textured content or structured noise, leaving a potentially nonzero interaction between surrogate residuals and prediction errors.

These examples connect existing assumptions to a common objective discrepancy. When the assumptions hold only approximately, the trace need not vanish, motivating explicit control of the remaining interaction. LoTA-N2N estimates this interaction with frozen-teacher proxies and penalizes regional absolute magnitudes to prevent cancellation. The next section derives the resulting supervised-risk bound, including the slack introduced by proxy error.

C Complete Method and Proofs

The main paper summarizes the two-stage pipeline. For completeness, given one noisy image $\mathbf{y}$, two complementary sub-images are generated by

$$\mathbf{y}_1 = \mathbf{k}_1 \otimes \mathbf{y}, \qquad \mathbf{y}_2 = \mathbf{k}_2 \otimes \mathbf{y}, \tag{27}$$

where $\mathbf{k}_1$ and $\mathbf{k}_2$ average the anti-diagonal and diagonal pixels of each non-overlapping 2×2 block, respectively. An unmatched final row or column is omitted when an image dimension is odd. Because these pairs need not satisfy N2N independence or clean-neighbor similarity, direct MSE training can retain a nonzero residual–prediction interaction.

C.1 Stage 1: MSE Pretraining

Both stages retain the baseline bidirectional pair loss and full-image consistency loss. Define $D_i(\mathbf{y}) = \mathbf{k}_i \otimes \mathbf{y}$ and the corresponding latent clean sub-images $\mathbf{x}_i = D_i(\mathbf{x})$. These clean sub-images are used only for analysis, not training. The baseline losses are defined as follows:

$$\mathcal{L}_{\mathrm{pair}}(\theta) = \frac{1}{2} \sum_{(i,j)\in\{(1,2),(2,1)\}} \mathbb{E}\|f_\theta(\mathbf{y}_i) - \mathbf{y}_j\|_F^2, \tag{28}$$

$$\mathcal{L}_{\mathrm{cons}}(\theta) = \frac{1}{2} \sum_{i=1}^{2} \mathbb{E}\|f_\theta(\mathbf{y}_i) - D_i(f_\theta(\mathbf{y}))\|_F^2, \tag{29}$$

$$\mathcal{L}_{\mathrm{pre}}(\theta) = \mathcal{L}_{\mathrm{MSE}}(\theta) = \mathcal{L}_{\mathrm{pair}}(\theta) + \mathcal{L}_{\mathrm{cons}}(\theta). \tag{30}$$

Gradients propagate through both student predictions in the consistency term. The trace decomposition concerns $\mathcal{L}_{\mathrm{pair}}$; the additional consistency term is nonnegative and remains part of the empirical baseline. Let θ_0 denote the pretrained parameters. Stage 2 initializes the student from this checkpoint; a separate frozen teacher provides fixed estimates of both clean sub-images:

$$\widehat{\mathbf{x}}_i = \mathrm{sg}\Big(\mathbf{k}_i \otimes \mathrm{clip}_{[0,1]}(f_{\theta_0}(\mathbf{y}))\Big), \quad i \in \{1, 2\}, \tag{31}$$

Here $\mathrm{sg}(\cdot)$ stops gradients through the frozen teacher predictions.

C.2 Reliability of the Estimated Trace

For direction $i \to j$, define $\mathbf{e}_i = \widehat{\mathbf{x}}_i - \mathbf{x}_i$, $\mathbf{p}_{\theta,i} = f_\theta(\mathbf{y}_i) - \mathbf{x}_i$, and $\mathbf{q}_{i\to j} = \mathbf{y}_j - \mathbf{x}_i$. Let

$$\begin{aligned} \mathcal{T}_{i\to j} &= \mathbb{E}\langle \mathbf{q}_{i\to j}, \mathbf{p}_{\theta,i} \rangle_F, \\ \widehat{\mathcal{T}}_{i\to j} &= \mathbb{E}\langle \mathbf{y}_j - \widehat{\mathbf{x}}_i, f_\theta(\mathbf{y}_i) - \widehat{\mathbf{x}}_i \rangle_F. \end{aligned} \tag{32}$$

Proposition 2 (Estimation error). If $\mathbb{E}\|\mathbf{e}_i\|_F^2 \le \epsilon^2$, $\mathbb{E}\|\mathbf{p}_{\theta,i}\|_F^2 \le P^2$, and $\mathbb{E}\|\mathbf{q}_{i\to j}\|_F^2 \le Q^2$, the estimation error satisfies the following bound:

$$|\widehat{\mathcal{T}}_{i\to j} - \mathcal{T}_{i\to j}| \le \epsilon(P+Q) + \epsilon^2. \tag{33}$$

Proof. Expanding with $\widehat{\mathbf{x}}_i = \mathbf{x}_i + \mathbf{e}_i$ gives

$$\widehat{\mathcal{T}} - \mathcal{T} = \mathbb{E}[\,-\langle \mathbf{e}_i, \mathbf{p}_{\theta,i}\rangle_F - \langle \mathbf{q}_{i\to j}, \mathbf{e}_i\rangle_F + \|\mathbf{e}_i\|_F^2]. \tag{34}$$

The result follows from the triangle and Cauchy–Schwarz inequalities.

Under the differentiation conditions of Proposition 1, if the denoiser Jacobian $\mathbf{J}_{\theta,i}$ satisfies $\|\mathbf{J}_{\theta,i}\|_{\mathrm{op}} \le K$, stop-gradient further gives

$$\|\nabla_\theta \widehat{\mathcal{T}}_{i\to j} - \nabla_\theta \mathcal{T}_{i\to j}\|_2 \le K\epsilon. \tag{35}$$

For fixed P and Q, a smaller teacher-error bound ϵ reduces the slack in the trace estimate. The gradient bound separately supports the diagnostic interpretation in Section D.6; it is not used as a training objective in the final method.

C.3 Local Low-Trace Adaptation

A global trace can be small even when large positive and negative regional interactions remain. Taking the absolute value within each region before aggregation removes this source of cancellation. We therefore partition each sub-image into M spatial regions $\{\Omega_m\}_{m=1}^M$ and assign area weights $w_m = |\Omega_m| / \sum_n |\Omega_n|$. Boundary regions may be smaller; equal-sized regions have $w_m = 1/M$. For direction $i \to j$, the true and estimated regional interactions are

$$\begin{aligned} \mathcal{T}_{i\to j}^{(m)} &= \mathbb{E}\big[\langle (\mathbf{q}_{i\to j})_{\Omega_m}, (\mathbf{p}_{\theta,i})_{\Omega_m}\rangle_F\big], \\ \widehat{\mathcal{T}}_{i\to j}^{(m)} &= \mathbb{E}\big[\langle (\mathbf{y}_j - \widehat{\mathbf{x}}_i)_{\Omega_m}, (f_\theta(\mathbf{y}_i) - \widehat{\mathbf{x}}_i)_{\Omega_m}\rangle_F\big]. \end{aligned} \tag{36}$$

Each inner product already averages over all scalar entries of its region; no additional division by region size is applied. The analysis uses expected interactions; training substitutes regional inner products computed from the single observed image. Complete image coverage gives $\mathcal{T}_{i\to j} = \sum_m w_m \mathcal{T}_{i\to j}^{(m)}$, with the analogous identity for the estimated interaction. The resulting bidirectional loss penalizes the absolute magnitude of each regional interaction:

$$\mathcal{L}_{\mathrm{LTrC}} = \frac{1}{2} \sum_{(i,j)\in\{(1,2),(2,1)\}} \sum_{m=1}^{M} w_m |\widehat{\mathcal{T}}_{i\to j}^{(m)}|. \tag{37}$$

Proposition 3 (Cancellation control). For either direction,

$$\left| \sum_m w_m \widehat{\mathcal{T}}_{i\to j}^{(m)} \right| \le \sum_m w_m |\widehat{\mathcal{T}}_{i\to j}^{(m)}|. \tag{38}$$

Therefore, the local loss upper-bounds the absolute global trace estimate and prevents cancellation between regions. Cancellation can still occur within individual regions.

To connect this value-level control directly to the inaccessible supervised objective, define

$$\begin{aligned} \mathcal{L}_{\mathrm{MSE}} &= \mathcal{L}_{\mathrm{pair}} + \mathcal{L}_{\mathrm{cons}}, \\ \mathcal{L}_{\mathrm{sup}}^{\mathrm{bi}} &= \frac{1}{2} \sum_{i\in\{1,2\}} \mathbb{E}\|f_\theta(\mathbf{y}_i) - \mathbf{x}_i\|_F^2, \end{aligned} \tag{39}$$

with the corresponding average residual-energy constant

$$C_{\mathrm{bi}} = \frac{1}{2} \sum_{(i,j)\in\{(1,2),(2,1)\}} \mathbb{E}\|\mathbf{y}_j - \mathbf{x}_i\|_F^2. \tag{40}$$

We also define the inaccessible true local-trace magnitude and its estimation error:

$$\begin{aligned}\mathcal{L}^{\star}_{\mathrm{LTrC}} &= \frac{1}{2}\sum_{(i,j)\in\{(1,2),(2,1)\}}\sum_{m=1}^{M} w_m|\mathcal{T}^{(m)}_{i\to j}|,\\ \Delta_{\mathcal{T}} &= \frac{1}{2}\sum_{(i,j)\in\{(1,2),(2,1)\}}\sum_{m=1}^{M} w_m|\widehat{\mathcal{T}}^{(m)}_{i\to j}-\mathcal{T}^{(m)}_{i\to j}|.\end{aligned} \tag{41}$$

Theorem 1 (Supervised-risk upper bound). For the bidirectional sub-image objective,

$$\begin{aligned}\mathcal{L}^{\mathrm{bi}}_{\mathrm{sup}}(\theta) &\le \mathcal{L}_{\mathrm{MSE}}(\theta) - C_{\mathrm{bi}} + 2\mathcal{L}^{\star}_{\mathrm{LTrC}}(\theta)\\ &\le \mathcal{L}_{\mathrm{MSE}}(\theta) - C_{\mathrm{bi}} + 2\mathcal{L}_{\mathrm{LTrC}}(\theta) + 2\Delta_{\mathcal{T}}(\theta).\end{aligned} \tag{42}$$

If the assumptions of Proposition 2 hold patch-wise with common bounds ϵ, P, and Q, then

$$\Delta_{\mathcal{T}} \le \epsilon(P+Q) + \epsilon^2. \tag{43}$$

Proof. Apply the exact decomposition to each direction:

$$\mathcal{L}^{i}_{\mathrm{sup}} = \mathcal{L}^{i\to j}_{\mathrm{ss}} - C_{i\to j} + 2\sum_{m} w_m \mathcal{T}^{(m)}_{i\to j}. \tag{44}$$

The first inequality follows from $\sum_m w_m\mathcal{T}^{(m)} \le \sum_m w_m|\mathcal{T}^{(m)}|$, averaging the two directions, and adding $\mathcal{L}_{\mathrm{cons}} \ge 0$. The second follows from $|\mathcal{T}^{(m)}| \le |\widehat{\mathcal{T}}^{(m)}| + |\widehat{\mathcal{T}}^{(m)} - \mathcal{T}^{(m)}|$. Averaging the regional bounds from Proposition 2 proves the last inequality.

Theorem 1 bounds expected supervised sub-image risk using reconstruction MSE, local trace magnitudes, and teacher-estimation error. This motivates a practical adaptation loss using only first-order derivatives. The coefficient 2 belongs to the population bound; the empirical loss instead uses a scheduled weight λ_{v} and within-image interaction estimates.

Proposition 1 remains important for interpretation: without additional regularity, no value-only penalty can universally guarantee gradient equality. The theorem concerns expected sub-image risk and does not establish an equivalent bound for full-image inference. Gradient discrepancy is evaluated only through the diagnostics in Section D.6.

The final fine-tuning objective is therefore

$$\mathcal{L}_{\mathrm{ft}} = \mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{v}}\mathcal{L}_{\mathrm{LTrC}}. \tag{45}$$

The trace weight follows a cosine schedule. The residual parameterization expresses the denoised output as $f_\theta(\mathbf{y}) = \mathbf{y} - r_\theta(\mathbf{y})$, where r_θ predicts the noise residual.

D Complete Experimental Protocol and Results

D.1 Datasets and Evaluation Protocol

We first evaluate controlled synthetic corruption on Kodak24, McMaster18 Zhang et al. (2011), Set14 Zeyde et al. (2010), and Set5 Bevilacqua et al. (2012). Cross-domain evaluation uses confocal images from FMD Zhang et al. (2019) and pediatric chest X-ray images Kermany et al. (2018). Kodak24, McMaster18, and confocal images are evaluated at 500×500, while the selected X-ray images are evaluated at 800×800. Each image is adapted independently without clean targets.

The standard synthetic evaluation includes additive Gaussian, signal-dependent Poisson, and mixed Poisson–Gaussian noise. The main benchmark, paired statistics, and Kodak gain-distribution figure use Gaussian read noise $r = 2$ on the 0–255 scale. The introductory McMaster18 matched-control plot and backbone-transfer study instead use Poisson peak 50 and Gaussian $\sigma = 20$. Section D.5 specifies a third mixed-noise setting and spatially varying Gaussian noise; Table 4 also includes a correlated-Gaussian stress check. Comparisons are made only within each noise setting. Inputs lie in $[0, 1]$ and are clipped after corruption. We evaluate PSNR and SSIM Wang et al. (2004).

Method	Confocal (500×500)				X-ray (800×800)			
	Gaussian noise		Poisson noise		Gaussian noise		Poisson noise	
	$\sigma = 5$	$\sigma = 10$	$\lambda = 60$	$\lambda = 50$	$\sigma = 5$	$\sigma = 10$	$\lambda = 60$	$\lambda = 50$
CBM3D	42.47	38.28	36.87	36.70	41.30	38.60	35.57	35.25
DIP2000	38.98	37.11	37.20	36.84	36.21	35.95	35.53	35.33
Noise2Fast	41.49	38.98	38.52	38.25	40.83	38.40	35.32	34.84
ZSN2N	44.13	39.01	39.81	39.33	42.04	39.06	35.79	35.31
LoTA-N2N	**44.21**	**39.26**	**40.17**	**39.65**	**42.96**	**39.74**	**35.81**	**35.35**

Table 6: Quantitative comparison of different denoising methods on the Confocal and X-ray datasets. The evaluation metric is PSNR (dB), and the best result in each setting is highlighted in bold. The LoTA-N2N row uses the final floor-0.15 configuration.

Aggregation protocols are stated with each table. In the selected noise-shift experiment, every cell averages five independently generated noise realizations per image using a fixed optimization seed of 2027, and all methods receive the same saved noisy inputs. This table reports mean metrics only; image-level bootstrap intervals for the main benchmarks are provided separately in Section D.4 under its three-noise-level aggregation protocol.

D.2 Implementation Details

The denoising backbone is a three-layer residual CNN with 48 hidden channels, 3×3, 3×3, and 1×1 convolution kernels, and Leaky-ReLU after the first two convolutions. Stage 1 minimizes the bidirectional pair MSE and full-image consistency in Eq. (30) for 3,000 Adam steps. The initial learning rate is 10^{-3} and halves every 1,000 steps. The resulting checkpoint initializes the student and a separate frozen teacher.

Stage 2 restarts Adam and adapts the student for 800 steps, with a cosine learning-rate schedule from 10^{-4} to 10^{-5}. The teacher remains fixed. The trace weight starts at $\lambda_{\mathrm{v}} = 8$ and follows a cosine decay to 15% of this value, giving a final weight of 1.2. Unless stated otherwise, all LoTA-N2N results use this floor-0.15 configuration. Both stages retain the pair and consistency losses; Stage 2 additionally penalizes the local trace magnitude.

Regions have width 32 on the sub-image grid. LTrC uses every region at each update, with area weights accounting for smaller boundary regions. The `sampled_regions` field applies only to gradient-trace ablations; LTrC is the only trace regularizer in the final objective, avoiding double backpropagation. Appendix D.7 documents configuration selection and the subsequent sensitivity study on Set12 images.

D.3 Additional Comparisons and Cross-Domain Results

For fairness, all rerun zero-shot methods use the same noisy inputs, image range, clipping policy, and metric implementation. N2N is trained for 100 epochs on Kodak24 and 250 epochs on McMaster18, while DIP is run for at most 2,000 iterations. We use author-provided single-image implementations when available, retaining published defaults apart from the shared image-range conversion.

To evaluate transfer beyond natural images, we apply the same zero-shot adaptation protocol to confocal microscopy and pediatric chest X-ray images, without external domain-specific training data. Table 6 reports PSNR under controlled Gaussian and Poisson corruptions, testing the objective on images with different textures, contrasts, and anatomical structures.

LoTA-N2N achieves the highest PSNR in all eight confocal and X-ray settings. Compared with ZSN2N, the gains are 0.08–0.36 dB on confocal images and 0.02–0.92 dB on X-ray images. The improvement is especially clear for Gaussian-corrupted X-ray images. These results support local trace correction across image domains under the tested corruptions.

Dataset	Noise family	Matched baseline	ΔPSNR (dB), 95% CI	Win rate (%)
Kodak24	Gaussian	ZSN2N-3800*	$+0.37$ $[+0.30, +0.44]$	91.7
Kodak24	Poisson	ZSN2N-3800*	$+0.39$ $[+0.31, +0.47]$	95.8
Kodak24	Mixed ($r = 2$)	ZSN2N-3800*	$+0.39$ $[+0.31, +0.47]$	95.8
McMaster18	Gaussian	ZSN2N-3800*	$+0.50$ $[+0.40, +0.61]$	100.0
McMaster18	Poisson	ZSN2N-3800*	$+0.58$ $[+0.47, +0.69]$	100.0
McMaster18	Mixed ($r = 2$)	ZSN2N-3800*	$+0.58$ $[+0.47, +0.70]$	100.0
Set14	Gaussian	ZSN2N-3800*	$+0.19$ $[-0.004, +0.37]$	71.4
Set14	Poisson	ZSN2N-3800*	$+0.31$ $[+0.13, +0.47]$	85.7
Set14	Mixed ($r = 2$)	ZSN2N-3800*	$+0.31$ $[+0.13, +0.47]$	85.7

Table 7: Paired comparison with the iteration-matched ZSN2N-3800 control. For each image, PSNR differences are averaged over the three noise levels before 10,000 image-level bootstrap resamples. ZSN2N-3800* starts from the same 3,000-step checkpoint and performs 800 additional MSE-only steps with restarted Adam and the Stage 2 learning-rate schedule. LoTA-N2N uses its final floor-0.15 configuration for all reported noise families.

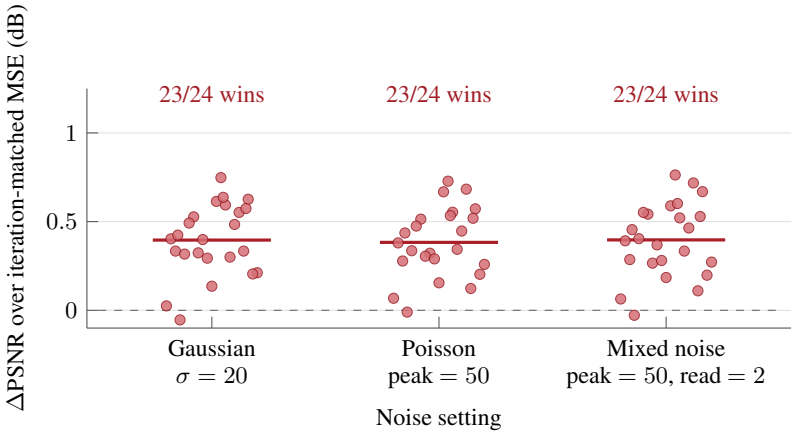


Figure 4: Per-image Stage 2 gain of LoTA-N2N over an iteration-matched MSE control using the standard three-layer CNN on Kodak24. Both methods start from the same 3,000-step Stage 1 checkpoint and run 800 additional Stage 2 optimization steps; LoTA-N2N uses the final trace-weight floor of 0.15. Points represent individual images; horizontal bars mark the average gain across images.

D.4 Statistical Reliability

Average PSNR can conceal whether an improvement is broadly distributed or driven by a few favorable images. We therefore compare LoTA-N2N with the iteration-matched ZSN2N-3800 MSE control on identical noisy inputs. Table 7 reports the paired mean PSNR difference, its 95% bootstrap confidence interval, and the per-image win rate for each noise family.

Figure 4 complements the three-level bootstrap summary with per-image gains at one representative level from each noise family.

The Kodak24 and McMaster18 confidence intervals exclude zero for all reported noise families, with win rates of at least 91.7% and 100%, respectively. Set14 also shows positive confidence intervals for Poisson and mixed noise. At the representative Kodak24 levels in Figure 4, LoTA-N2N improves 23 of 24 images in each setting. These per-image gains are computed at individual noise levels, whereas the table averages three levels before bootstrapping.

D.5 Robustness under Noise-Distribution Shift

We examine whether explicit local trace control remains useful across changes in noise level and distribution. We evaluate three controlled conditions on Kodak24: IID Gaussian noise with $\sigma = 25/255$, spatially varying Gaussian noise with a horizontal standard-deviation map from $5/255$ to

Method	IID Gaussian	Spatially varying	Poisson–Gaussian
Self2Self†	27.61/0.7762	27.21/0.7657	27.13/**0.7711**
Noise2Fast	27.65/0.7523	25.24/0.6880	26.06/0.7002
ZSN2N-3800*	29.20/0.7775	28.09/0.7583	27.96/0.7495
Global TrC	29.55/0.7942	28.35/0.7632	28.31/0.7655
LoTA-N2N	**29.62/0.7975**	**28.40/0.7680**	**28.37**/0.7702

Table 8: Selected non-correlated noise evaluations on Kodak24. Each cell reports mean PSNR (dB)/SSIM over all 24 images and five matched noise realizations. All optimization-based methods use approximately 3,800 optimization steps. *The MSE control and LoTA-N2N share the same 3,000-step Stage 1 checkpoint and run 800 additional steps with restarted Adam and the same Stage 2 learning-rate schedule. LoTA-N2N and Global TrC use the final floor-0.15 configuration. †Self2Self uses a partial-convolution reimplementation with 100 stochastic predictions. Boldface marks the best value for each metric.

$50/255$, and mixed Poisson–Gaussian noise with peak 30 followed by Gaussian read noise with $\sigma = 5/255$. No method is given the spatial noise map or the mixed-noise parameters. Table 8 reports the resulting PSNR/SSIM for all evaluated methods.

Table 8 shows that LoTA-N2N obtains the best PSNR in all three conditions and the best SSIM under IID and spatially varying Gaussian noise. Relative to the checkpoint-matched ZSN2N-3800* control, the PSNR gains are 0.42, 0.31, and 0.41 dB, respectively. Local correction also improves on Global TrC by 0.07, 0.05, and 0.06 dB, with higher SSIM in each setting. These matched comparisons support the benefit of controlling regional interactions under the evaluated noise shifts.

D.6 Mechanism Validation

Theorem 1 bounds expected supervised sub-image risk using reconstruction, local trace magnitude, and estimation error. We additionally use synthetic data with known clean images to test the first-order behavior characterized by Proposition 1. These diagnostic quantities are never exposed to the final training loss. At fixed checkpoints, let $\mathbf{g}_{\mathrm{sup}}$ and $\mathbf{g}_{\mathrm{ss}}$ denote gradients of the supervised and self-supervised pair-MSE objectives with respect to all trainable parameters; the consistency term is excluded from both diagnostics. We report the normalized gradient error

$$G_{\mathrm{rel}} = \frac{\|\mathbf{g}_{\mathrm{ss}} - \mathbf{g}_{\mathrm{sup}}\|_2}{\|\mathbf{g}_{\mathrm{sup}}\|_2 + \delta}, \tag{46}$$

and their cosine similarity

$$G_{\cos} = \frac{\langle \mathbf{g}_{\mathrm{ss}}, \mathbf{g}_{\mathrm{sup}} \rangle}{\|\mathbf{g}_{\mathrm{ss}}\|_2 \|\mathbf{g}_{\mathrm{sup}}\|_2 + \delta}. \tag{47}$$

Global TrC averages the absolute global traces of the two directions; LTrC instead takes absolute values before aggregating regions. We quantify regional cancellation using the following ratio:

$$R_{\mathrm{cancel}} = 1 - \frac{\frac{1}{2}\sum_{(i,j)} |\sum_m w_m \widehat{\mathcal{T}}^{(m)}_{i\to j}|}{\mathcal{L}_{\mathrm{LTrC}} + \delta}. \tag{48}$$

A value near one indicates that regional estimates largely cancel relative to their absolute magnitudes. The ratio describes relative cancellation, not the absolute magnitude of local interactions.

For the oracle diagnostic, let $\overline{\mathcal{T}}$ and $\overline{\widehat{\mathcal{T}}}$ be the signed averages of the two directional global traces. The per-image relative estimation error is

$$E_{\mathcal{T}} = \frac{|\overline{\widehat{\mathcal{T}}} - \overline{\mathcal{T}}|}{|\overline{\mathcal{T}}| + \delta}, \qquad \delta = 10^{-12}. \tag{49}$$

Table 2 averages this ratio over images; it is not computed from aggregate trace magnitudes. The risk bound accounts for teacher error in connecting proxy estimates to supervised reconstruction.

For ablation only, we also instantiate the direct gradient-trace penalty

$$\mathcal{L}_{\mathrm{GTrC}}^{i\to j} = \frac{1}{M}\sum_{m=1}^{M} \|\nabla_\theta (M w_m \widehat{\mathcal{T}}_{i\to j}^{(m)})\|_2^2,$$
$$\mathcal{L}_{\mathrm{GTrC}} = \frac{1}{2}\sum_{(i,j)\in\{(1,2),(2,1)\}} \mathcal{L}_{\mathrm{GTrC}}^{i\to j}. \tag{50}$$

The implementation samples one area-scaled regional value per direction for this gradient-only ablation. LTrC itself uses all regional values and requires no double backpropagation during the final training procedure. Table 2 compares these penalties with identical adaptation budgets.

On McMaster18 at $\sigma = 10$, LTrC reaches 34.58 dB, compared with 34.17 dB for matched MSE adaptation and 34.49 dB for Global TrC. Its estimated local interaction is also smaller: 2.02×10^{-5} versus 5.14×10^{-5} and 3.26×10^{-5}, respectively. This association supports the motivation for controlling regional interactions rather than only their global sum.

Figure 5 complements these averages with per-image results and comparisons across the four adaptation objectives. LTrC improves PSNR over matched MSE adaptation on all 18 images while reducing the estimated local trace. Across objectives, lower local-trace magnitude and gradient error are associated with larger PSNR gains. These are descriptive diagnostics, not a guarantee that reducing trace magnitude also aligns optimization gradients.

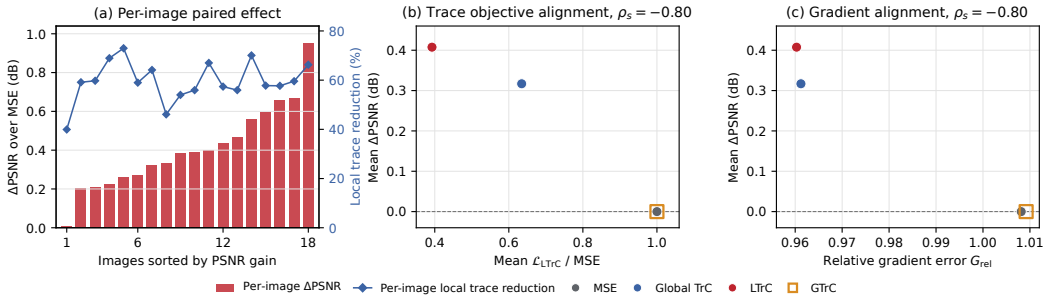


Figure 5: Additional mechanism diagnostics on McMaster18 at Gaussian $\sigma = 10$. (a) Per-image PSNR gains and estimated local-trace reductions of LTrC relative to iteration-matched MSE adaptation, sorted by PSNR gain. (b) Mean PSNR gains versus local-trace magnitude normalized by the MSE control. (c) Mean PSNR gains versus normalized gradient error G_{rel}. The correlations summarize the four plotted objectives; the final method does not optimize any of the plotted gradient quantities as part of the final training objective.

D.7 Additional Configuration and Computational Cost

The base configuration—region width 32, initial trace weight $\lambda_{\mathrm{v}} = 8$, 800 Stage 2 steps, and an original floor of 0.10—was selected on five Set12 images. The floor was subsequently increased to 0.15 after inspecting Kodak24/McMaster18 benchmark results. A later sensitivity study used Set12 images 01–05 at Gaussian levels $\sigma = 10$ and 25. This retrospective study assesses sensitivity rather than establishing an independent hyperparameter selection. Table 9 reports the computational cost of this configuration across image resolutions.

At 500×500, LoTA-N2N takes 20.37 seconds per image compared with 20.24 seconds for ZSN2N, with the same reported peak memory of 0.28 GB (Table 9). LTrC adds no trainable parameters: both methods use the 0.022M-parameter backbone. The final method thus provides local interaction control with a small measured runtime increment at this resolution and without the additional double-backpropagation cost incurred by GTrC.

Method	Steps/image	Params (M)	125×125		250×250		500×500	
			Time (s)	Mem. (GB)	Time (s)	Mem. (GB)	Time (s)	Mem. (GB)
Noise2Fast	≈ 3801	0.259	22.11	0.03	25.19	0.09	33.61	0.34
ZSN2N	3800	0.022	16.51	0.02	18.68	0.07	20.24	0.28
Self2Self[†]	3800	0.991	45.82	6.91	44.41	15.16	69.23	10.66
Global TrC	3800	0.022	18.19	0.02	18.12	0.07	19.85	0.28
LoTA-N2N	3800	0.022	18.81	0.02	18.63	0.07	20.37	0.28

Table 9: Resolution-scaling runtime and memory comparison on an NVIDIA H100 80GB HBM3 using random RGB inputs. Each entry reports the average over three random inputs. The LoTA-N2N timing uses the final floor-0.15 Stage 2 configuration. Noise2Fast is trained for 1267 steps per channel, corresponding to approximately 3801 total optimization steps per RGB image. PyTorch methods report peak `torch.cuda.max_memory_allocated`; Self2Self memory is measured by peak `nvidia-smi` sampling because TensorFlow compat-v1 does not expose the same allocator statistic. [†]Self2Self is measured with a TensorFlow compat-v1 reimplementation of the partial-convolution Self2Self runtime and includes 100 final stochastic predictions; its parameter count excludes the fixed buffer used to average predictions.

D.8 Teacher-Quality Analysis

We vary the Stage 1 checkpoint used to construct frozen clean-image proxies on Set12 under Gaussian, Poisson, and mixed noise. Every adaptation follows the same 800-step MSE+LTrC Stage 2 protocol with floor 0.15. Figure 6 reports teacher and adapted PSNR together with relative trace-estimation and trace-gradient errors computed using clean references only for evaluation. These references are never used during training.

Stronger Stage 1 checkpoints generally give higher adapted PSNR and lower trace-estimation error across the four reported settings. This pattern is consistent with the role of proxy accuracy in the estimation-error term of Theorem 1. The trace-gradient curves are not uniformly monotonic, reinforcing the distinction between estimating trace values accurately and controlling their first-order behavior. Together, these diagnostics characterize the reliability of the frozen proxies used during the final 800-step local-trace adaptation procedure.

D.9 Additional Design-Ablation Results

Figure 7 expands Table 3 by showing both final PSNR and the gain from each setting's own Stage 1 model. The bidirectional-consistency variants A_2 and A_4 show positive Stage 2 gains at all four Gaussian levels. Combining this training formulation with residual prediction (A_4) gives the highest final PSNR, exceeding the strongest alternative by 0.45, 0.36, 0.22, and 0.13 dB, respectively. Because bidirectionality and consistency are enabled together, this comparison does not isolate their individual contributions to reconstruction performance.

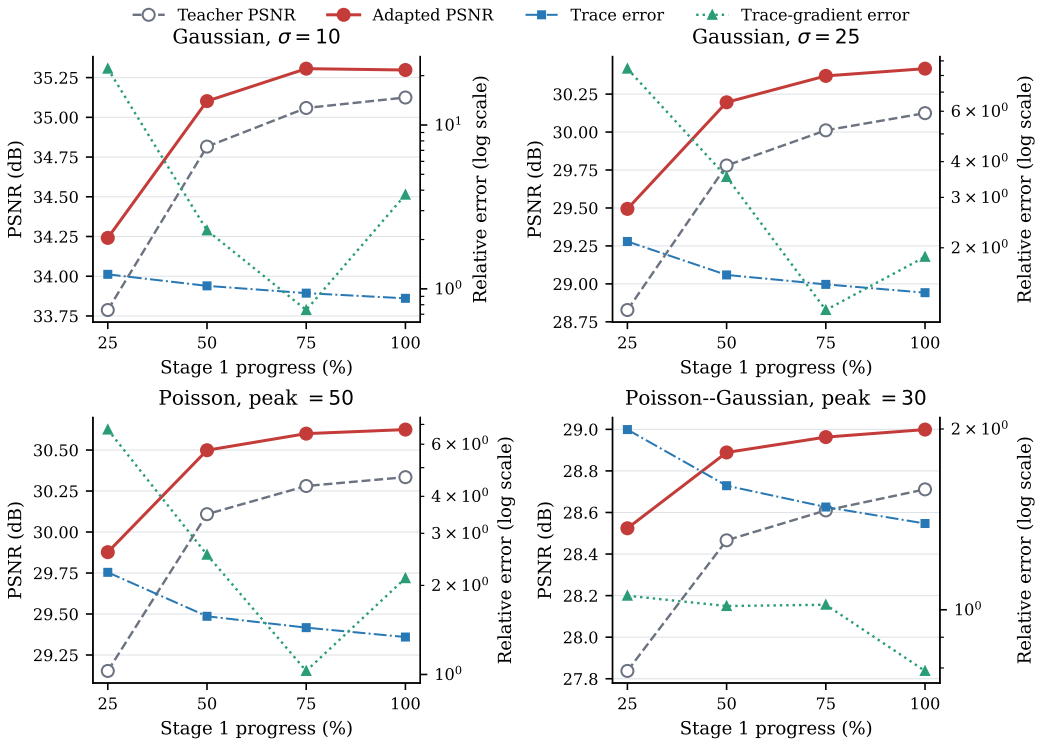


Figure 6: Effect of Stage 1 checkpoint quality on Set12. Each panel varies the checkpoint from 25% to 100% of Stage 1 training under the indicated corruption. Curves report teacher PSNR, adapted PSNR, relative trace-estimation error, and relative trace-gradient error; the error curves use the logarithmic right axis. All adapted models use 800 MSE+LTrC steps with floor 0.15. Trace-gradient error is evaluated only as a diagnostic and is excluded from the final training objective.

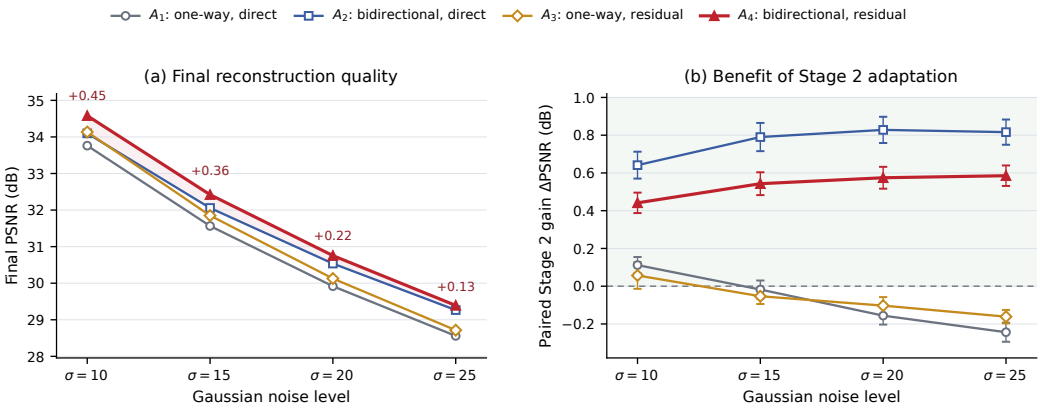


Figure 7: Design ablation on McMaster18 under the final floor-0.15 configuration. (a) Final PSNR, with annotations showing the margin of A_4 over the strongest alternative. (b) Paired Stage 2 PSNR gains with standard-error bars across images. The figure's bidirectional labels denote joint bidirectional pairing and full-image consistency, as specified in Table 3. Each setting has its own Stage 1 checkpoint and receives 800 additional adaptation steps. These checkpoints are separate from those used for the main benchmark results.

D.10 Supplementary Sensitivity Results

Following the configuration chronology in Appendix D.7, we report a retrospective one-factor sensitivity study on Set12 images 01–05 at Gaussian $\sigma = 10$ and 25. The default uses region width 32, initial trace weight $\lambda_v = 8$, floor 0.15, and 800 Stage 2 steps. Figure 8 varies one setting at a time and reports paired changes relative to this default, keeping the remaining settings fixed.

The default weight of 8 gives the highest PSNR among the three tested weights at both noise levels, and extending adaptation from 800 to 1,000 steps yields only a small additional PSNR gain. Lower normalized trace alone does not identify the best reconstruction setting. In particular, changing region width also changes the aggregation geometry, so trace magnitudes across widths must be interpreted together with PSNR rather than as directly comparable measures of correction strength within the evaluated adaptation protocol.

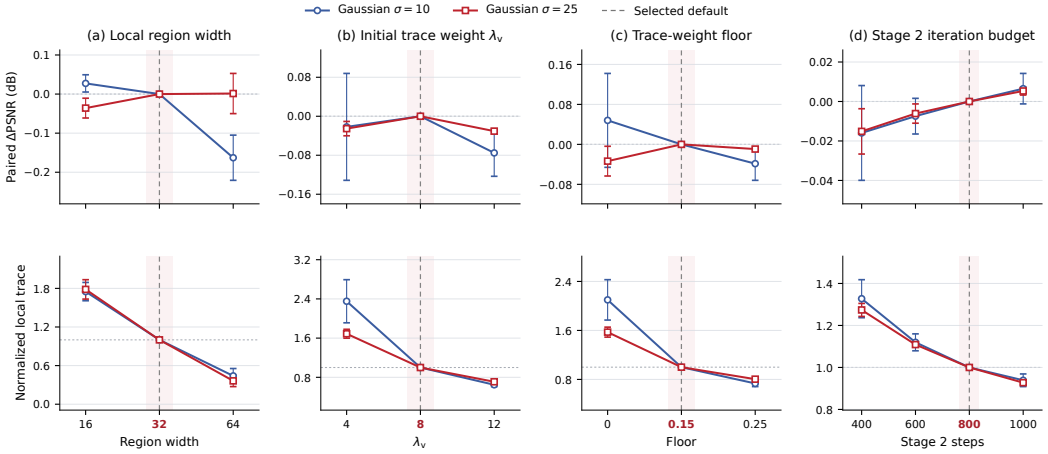


Figure 8: One-factor sensitivity on five Set12 images at Gaussian $\sigma = 10$ and 25. Columns vary region width, initial trace weight, trace-weight floor, and Stage 2 budget. The top row reports paired PSNR changes from the default; the bottom row reports local-trace magnitude normalized by each image's default value. Curves show five-image means with standard-error bars. Dashed lines and highlighted ticks mark the final settings assessed in this retrospective sensitivity study.

D.11 Additional Qualitative Results

Figure 9 shows confocal and chest-radiograph examples under synthetic corruption. LoTA-N2N reduces visible noise and achieves higher image-specific PSNR/SSIM than the shown baselines in these two illustrated cross-domain denoising examples.

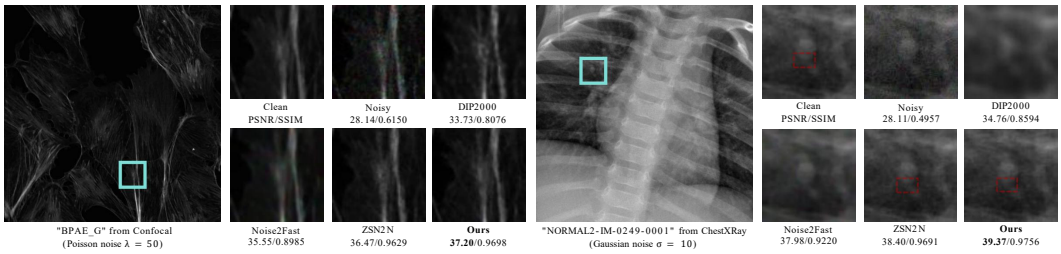


Figure 9: Additional qualitative examples from confocal microscopy (`BPAE_G`, Poisson peak 50) and chest radiography (`NORMAL2-IM-0249-0001`, Gaussian $\sigma = 10$). Cyan boxes mark the enlarged regions. The PSNR/SSIM labels are image-specific scores, rather than the dataset averages reported in the quantitative tables.

Figure 10 provides further natural-image comparisons on McMaster18, Set14, and Set5 under Poisson and mixed Poisson–Gaussian noise. The enlarged regions show fabrics, skin, flowers, and thin branches. In these examples, LoTA-N2N reduces visible residual noise while retaining prominent boundaries and structures. These selected examples complement the dataset-level comparisons rather than establishing performance across untested image content.

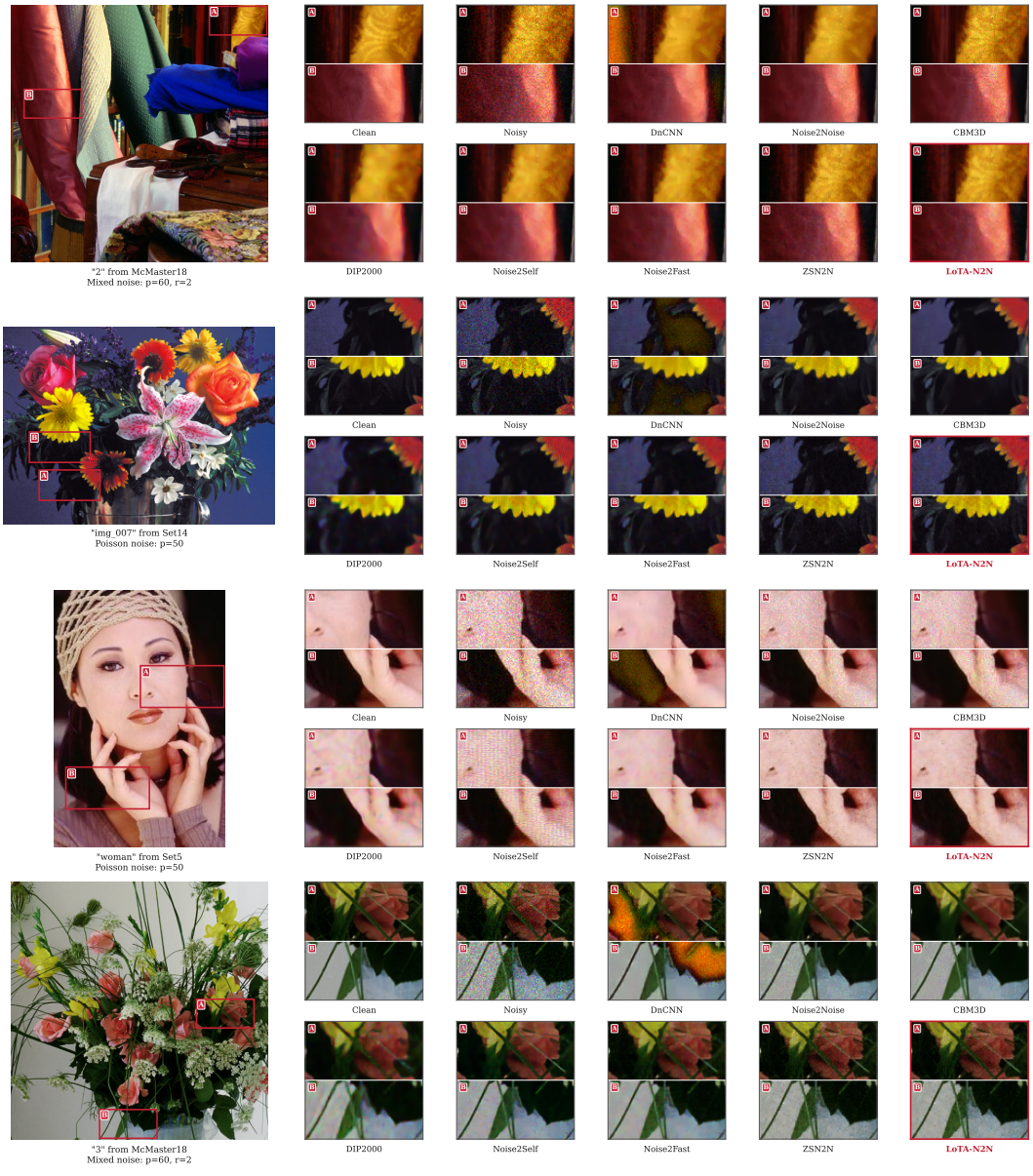


Figure 10: Additional natural-image comparisons. From top to bottom: McMaster18 image 2 under mixed noise (Poisson peak 60, Gaussian read noise $\sigma = 2$); Set14 image `img_007` and Set5 image `woman` under Poisson noise (peak 50); and McMaster18 image 3 under mixed noise (peak 60, $\sigma = 2$). Boxes identify the corresponding enlarged regions for all methods in each comparison.