PHASEFUSE: RECURSIVE FUSION OF ANALYTIC SOLVERS FOR SELF-SUPERVISED PHASE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Phase retrieval reconstructs an object from intensity-only measurements, but analytic solvers can respond differently to limited and noisy observations. We study whether these complementary estimates can be combined without clean reconstruction targets or natural-image training data. PhaseFuse aligns the global phases of five analytic solver outputs and of a candidate that keeps the current estimate, then uses a compact network to predict spatial fusion weights. The normalized blend is trained on procedural phase objects using independent noisy input and target measurements. At inference, a recursive schedule restarts the solvers from the fused estimate after short update blocks, reusing the single-pass-trained selector. On 200 BSDS500 test images encoded as 128×128 phase-only objects, with two coded diffraction patterns and 30 dB Gaussian measurement SNR, recursive pixel-wise fusion (PhaseFuse-PS) reaches 35.97 dB mean phase-PSNR. This is 7.27 dB above single-pass fusion and 9.32 dB above the strongest configured standalone expert, at 1.76 times the measured inference time of single-pass fusion, and every test image improves. Image-level selection (PhaseFuse-GH) reaches 36.45 dB under the same condition. Routing becomes more concentrated as the snapshot count increases. These results support further study of learned solver combination and repeated constraint enforcement under matched computation and on experimental imaging data.

1 INTRODUCTION

Phase retrieval (PR) recovers a complex-valued signal from intensity-only measurements (Shechtman et al., 2015). It is the computational core of coherent imaging modalities in which detectors record only the magnitude of a wavefield, including coherent diffraction imaging (Miao et al., 1999), ptychography (Thibault et al., 2008), and X-ray crystallography (Millane, 1990). Coded diffraction patterns (CDP) modulate the object with random masks before propagation, which improves identifiability under suitable mask and sampling assumptions (Candès et al., 2015); the measurements always remain invariant to a global phase. With few snapshots and noisy intensities, however, PR remains a nonconvex problem in which reconstruction algorithms stagnate, diverge, or converge to spurious solutions.

Analytic iterative solvers, based on alternating projections (Chen et al., 2018; Luke, 2005), gradient flows on intensity or amplitude losses (Chen & Candès, 2017; Wang et al., 2018), or robust splitting (Jiang et al., 2017), avoid a learned image prior, but their behavior depends on the sensing operator, initialization, data fidelity, and stopping rule, and no single solver is best across images, noise levels, and sampling conditions. Among learned methods, many supervised or pretrained-prior approaches, including end-to-end networks (Sinha et al., 2017), unrolled architectures (Jerez et al., 2024), and denoiser-based (Metzler et al., 2018) or diffusion priors (Chung et al., 2023), require clean targets or external image collections, which scientific imaging rarely provides. Untrained and measurement-supervised approaches, including PhysenNet (Wang et al., 2020), AutoPhaseNN (Yao et al., 2022), and PtychoPINN (Hoidn et al., 2023), avoid paired clean targets under their respective imaging models but learn or optimize the reconstruction itself, and hybrid pipelines such as I2I-PR (Kaya & Oktem, 2026) refine classical estimates with a learned image-to-image model. Learned spatial fusion of analytic estimators is established in image restoration (Hammond & Simoncelli, 2007), and phase-retrieval methods such as guided hybrid input-output (Chen et al., 2007), memetic phase retrieval (Colombo et al., 2017), and SPRING (Colombo et al., 2025) exchange information among

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.

reconstruction trajectories before convergence. We ask instead whether learning can be restricted to combining the outputs of fixed analytic solvers, trained from noisy measurements alone, without ground truth or natural images.

We answer this question with PhaseFuse, a self-supervised method that learns to combine the reconstructions of analytic PR solvers rather than select one or replace them. Its premise is that the solvers are complementary: a solver that diverges on one image, or in one region, is often accurate on another, so a combination that trusts each solver where it is reliable can exceed the best of them (Figure 1).

Given the measurements, PhaseFuse runs five analytic solvers, spanning projection, gradient, and robust splitting methods, from a shared spectral initialization, aligns their candidate reconstructions to a common global phase, and lets a compact convolutional network, informed by how well each candidate explains the measurements, assign every pixel a weight vector over the experts. The reconstruction is the weighted blend of the candidates, projected onto the unit circle, so the network learns only how much to trust each solver at each location. A spectral expert that returns its input unchanged, the shared initialization at the first decision and the current estimate afterwards, completes the library, and the same weights can also be read at the image level to commit to a single expert.

Learning such a combination would normally require ground truth. PhaseFuse instead observes each training object twice with independent noise, drives the solvers and the selector with the first observation, and scores the reconstruction by how well it predicts the second; in expectation this equals the error against the noiseless measurements plus a term that does not depend on the selector, which applies the Noise2Noise principle (Lehtinen et al., 2018) in the intensity domain. The training objects are procedurally generated from affine iterated function systems (Nakamura et al., 2023), so the selector sees neither ground-truth reconstructions nor natural images.

Beyond the design of the combination, its timing is decisive. A single-pass fusion runs every solver to completion and combines their outputs once, so the solvers never benefit from one another. Recursive expert fusion instead splits the same per-solver budget into short blocks and restarts all solvers from the fused estimate after each block, reusing the selector without retraining at the cost of one fusion step per block.

On all 200 BSDS500 test images (Martin et al., 2001; Arbeláez et al., 2011), with two snapshots and 30 dB SNR, recursive fusion reaches 35.97 dB with PhaseFuse-PS, 7.27 dB above single-pass fusion and 9.32 dB above the strongest individual solver at an identical budget of 200 updates per solver, and improves every image at $1.76\times$ the inference time of a single pass. Image-level selection is ahead of the per-pixel blend at high SNR and on large images and behind it at low SNR and on small images, the learned weights concentrate on a single expert as the snapshot count grows, and under recursion a selector trained only on procedural objects matches or exceeds one trained on natural images.

Our contributions are:

- A phase-aware fusion architecture that predicts spatial weights over heterogeneous analytic PR candidates and projects their weighted sum onto the unit circle.
- A self-supervised training scheme with a measurement-domain noisy-target objective on procedural objects, without clean targets or natural images.
- Recursive expert fusion, an inference schedule that reuses the single-pass-trained selector and improves on single-pass fusion by 7.27 dB and on the strongest expert by 9.32 dB at the same per-solver budget.
- An analysis of per-pixel versus image-level combination and of solver budget, schedule, training length, and training data.

2 Phase Retrieval Inverse Problem

We consider phase retrieval (PR) under the coded diffraction pattern (CDP) model. Let $\boldsymbol{x} \in \mathbb{C}^n$ be the unknown signal and $\{\boldsymbol{m}_\ell\}_{\ell=1}^{L} \subset \mathbb{C}^n$ known masks with unit magnitude and independent, uniformly distributed phases. With $\boldsymbol{F}$ the unitary two-dimensional discrete Fourier transform acting on vectorized images, each mask defines the operator $\boldsymbol{A}_\ell = L^{-1/2}\boldsymbol{F}\,\mathrm{diag}(\boldsymbol{m}_\ell)$, and stacking the L operators gives $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, with $m = Ln$ and $\boldsymbol{A}^H\boldsymbol{A} = \boldsymbol{I}_n$. Denoting by $\boldsymbol{a}_i^H$ the i-th row of $\boldsymbol{A}$, the

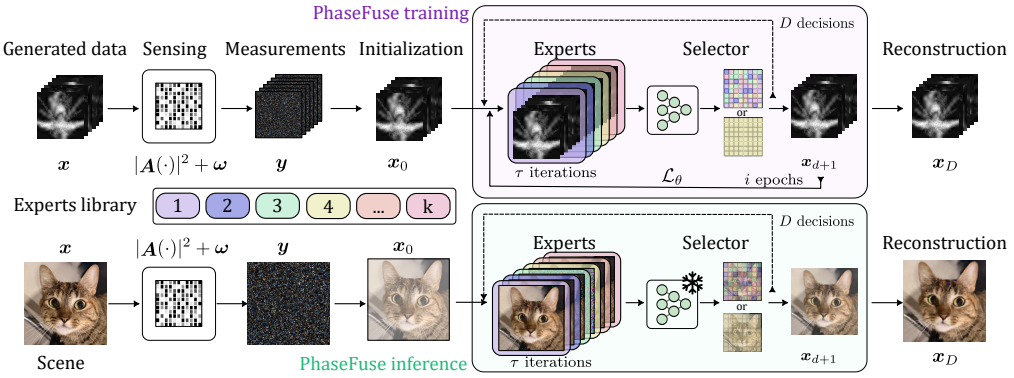


Figure 1: Overview of PhaseFuse. Top: training on generated objects. The experts start from $\boldsymbol{x}_0$ and run τ iterations; the selector turns their phase-aligned candidates into per-pixel weights (PhaseFuse-PS, patterned grid) or one expert per image (PhaseFuse-GH, uniform grid), and $\boldsymbol{x}_{d+1}$ restarts the experts for D decisions. The loss $\mathcal{L}_\theta$, computed against a second independent noisy observation of the object (Section 3.4), updates the selector over i training epochs; the main selectors use PhaseFuse-PS with $D = 1$, and $D > 1$ fits the selector to the recursive schedule (Section 4.3). Bottom: inference with the frozen selector.

measurements are

$$y_i = |\boldsymbol{a}_i^H \boldsymbol{x}|^2 + \omega_i, \quad i = 1, \dots, m, \qquad \text{or} \qquad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2 + \boldsymbol{\omega}, \tag{1}$$

where $\boldsymbol{\omega}$ is additive noise and the magnitude and square act element-wise. The measurements are invariant to a global phase factor, since $\boldsymbol{x}$ and $e^{j\phi}\boldsymbol{x}$ produce identical measurements for any $\phi \in [0, 2\pi)$, so the signal can only be recovered up to a global phase. In this work, $\boldsymbol{x}$ is a phase-only object, so under Gaussian noise PR is posed as

$$\min_{\boldsymbol{x} \in \mathbb{C}^n} \ \tfrac{1}{2} \big\| \boldsymbol{y} - |\boldsymbol{A}\boldsymbol{x}|^2 \big\|_2^2 \quad \text{subject to} \quad |x(p)| = 1 \text{ for every pixel } p, \tag{2}$$

a nonlinear and nonconvex problem. The analytic experts of Section 3.2 solve unconstrained versions of it, and the fusion imposes the constraint (Section 3.3).

3 Method

3.1 Procedural image generation

We train only on procedurally generated images, which carry no semantic content, cannot overlap with the evaluation images, and provide diverse local structure, such as edges, fine textures, and smooth regions at multiple scales. We build on the one-instance FractalDB construction of Nakamura et al. (2023), in which each category is an affine iterated function system (IFS) $\{(\boldsymbol{B}_j, \boldsymbol{b}_j)\}_{j=1}^{J}$, with $\boldsymbol{B}_j \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{b}_j \in \mathbb{R}^2$. We generate a finite random orbit $\boldsymbol{q}_{t+1} = \boldsymbol{B}_{j_t}\boldsymbol{q}_t + \boldsymbol{b}_{j_t}$, with $\Pr(j_t = j) = |\det \boldsymbol{B}_j| / \sum_k |\det \boldsymbol{B}_k|$ so that each map is visited in proportion to the area it covers; because random affine maps need not be contractive, only systems whose orbits stay bounded are kept (Appendix A.1), and the orbit is rasterized and augmented in image space. To make structures of different scales coexist, two augmentations $I^{(a)}, I^{(b)}$ of the same category are combined as

$$I^{\text{synth}} = \mathcal{N}\big[(1-\alpha)\,\mathcal{U}\mathcal{D}(I^{(a)}) + \alpha\, I^{(b)}\big], \qquad \alpha \sim p(\alpha), \tag{3}$$

where $\mathcal{D}$ and $\mathcal{U}$ are downsampling and upsampling operators and $\mathcal{N}$ is min–max normalization; the first term contributes coarse structure and the second fine detail. Implementation details are given in Appendix A.1.

3.2 Expert library

The method builds on a fixed library of K analytic PR solvers, or experts, indexed by $\mathcal{E} = \{1, \dots, K\}$. Each expert e applies t updates of its own rule from an initial estimate $\boldsymbol{x}_0$ and returns $\boldsymbol{v}_e =$

$\mathcal{F}_e^{(t)}(\boldsymbol{y}, \boldsymbol{A}; \boldsymbol{x}_0)$. All experts share the same measurements, sensing matrix, and initialization, so their differences reflect only the reconstruction mechanism; their rules and hyperparameters are fixed, and none uses a learned prior. Because each expert is defined from an arbitrary starting point, it can also be restarted from any intermediate estimate (Section 3.5). The shared initialization is the filtered spectral method of Jerez et al. (2024), which computes $\boldsymbol{x}_0$ by power iterations on $\frac{1}{m}\sum_i \mathcal{T}(y_i)\,\boldsymbol{a}_i\boldsymbol{a}_i^H$, where $\mathcal{T}$ retains the largest half of the intensities, with a low-pass filter applied after every multiplication (Appendix A.2); because of the filter, the result is not in general the leading eigenvector of that matrix.

The library represents the main families of analytic PR algorithms. With $\boldsymbol{b} = \sqrt{\max(\boldsymbol{y}, 0)}$, since additive noise can make intensities negative, projection experts seek the intersection of the measurement set $\mathcal{M} = \{\boldsymbol{u} : |\boldsymbol{u}| = \boldsymbol{b}\}$ and the range $\mathcal{R}$ of $\boldsymbol{A}$, with projections $P_{\mathcal{M}}(\boldsymbol{u}) = \boldsymbol{b} \odot \mathrm{ph}(\boldsymbol{u})$, where ph is the stabilized unit phasor of Appendix A.2, and $P_{\mathcal{R}} = \boldsymbol{A}\boldsymbol{A}^H$: the parallel alternating projection (PAP) method (Chen et al., 2018) alternates between them, and relaxed averaged alternating reflections (RAAR) (Luke, 2005) combines their reflections to escape stagnation. Gradient experts descend a data-fidelity loss: truncated Wirtinger flow (TWF) (Chen & Candès, 2017) follows the truncated gradient of a Poisson log-likelihood of the intensities and discards measurements with unreliable residuals, and reweighted amplitude flow (RAF) (Wang et al., 2018) uses the amplitude loss $\sum_i (|\boldsymbol{a}_i^H\boldsymbol{x}| - b_i)^2$ with adaptive weights. The robust splitting expert LAD–ADMM (Jiang et al., 2017) minimizes the least absolute deviation $\sum_i \big||\boldsymbol{a}_i^H\boldsymbol{x}| - b_i\big|$, which is less sensitive to large noise values. Finally, a spectral expert returns its input unchanged, which is the shared spectral initialization when it runs on its own and the current estimate within the recursion (Section 3.5); it lets the fusion keep that estimate, for a total of $K = 6$ experts.

Because the experts rely on different objectives and robustness mechanisms, their accuracy varies across images, noise levels, and spatial regions. Each expert, however, recovers the signal only up to its own global phase (Section 2), so the estimates must be aligned before they are combined.

3.3 Selector and fusion

Let $\mathcal{P}_{\mathbb{S}^1}(\boldsymbol{z}) = \boldsymbol{z}/|\boldsymbol{z}|$ denote the element-wise projection onto the unit circle. Because the fusion combines candidates through a complex weighted sum, candidates with different global phases may cancel each other even when all of them are accurate. Each candidate is therefore aligned to the reference $\boldsymbol{r} = \mathcal{P}_{\mathbb{S}^1}(\boldsymbol{x}_0)$,

$$\boldsymbol{u}_e = \mathcal{P}_{\mathbb{S}^1}(\boldsymbol{v}_e), \qquad \widetilde{\boldsymbol{v}}_e = \boldsymbol{u}_e \exp(-j \arg\langle \boldsymbol{r}, \boldsymbol{u}_e\rangle)\,, \tag{4}$$

where $\langle \boldsymbol{r}, \boldsymbol{u}\rangle = \boldsymbol{r}^H\boldsymbol{u}$. The aligned candidate does not depend on the global phase of $\boldsymbol{v}_e$ whenever $\langle \boldsymbol{r}, \boldsymbol{u}_e\rangle \neq 0$, so the selector is invariant to the global phase of each expert, while the reconstruction is equivariant to a common rotation of all inputs (Appendix A.3).

The selector g_θ is a convolutional encoder–decoder (Ronneberger et al., 2015) that receives, at each pixel p, the real and imaginary parts of $\widetilde{v}_e(p)\overline{r(p)}$, which describe the local deviation of each candidate from the reference; the real and imaginary parts of $r(p+e)\overline{r(p)}$ for the horizontal and vertical neighbors $p + e$ of each pixel, which describe the local structure of the reference; and the normalized residual $\rho_e = \||\boldsymbol{A}\boldsymbol{u}_e|^2 - \boldsymbol{y}\|_2/\|\boldsymbol{y}\|_2$ of each expert, broadcast over all pixels, which indicates how well each candidate explains the measurements without requiring ground truth. It outputs one logit $z_e(p)$ per expert and pixel, and a softmax over experts with temperature T turns them into weights $w_e(p)$ on the probability simplex. The per-pixel rule, PhaseFuse-PS, blends the aligned candidates as

$$\widehat{x}(p) = \mathcal{P}_{\mathbb{S}^1}\Big[\textstyle\sum_{e\in\mathcal{E}} w_e(p)\,\widetilde{v}_e(p)\Big], \qquad w_e(p) = \tfrac{\exp\{z_e(p)/T\}}{\sum_{k\in\mathcal{E}}\exp\{z_k(p)/T\}}, \tag{5}$$

while the global rule, PhaseFuse-GH, averages the logits over the image and commits to a single expert, $\widehat{\boldsymbol{x}} = \widetilde{\boldsymbol{v}}_{e^\star}$ with $e^\star = \arg\max_e n^{-1}\sum_p z_e(p)$. Both rules read the same logits. The blend is differentiable in θ wherever the weighted sum is nonzero (Appendix A.3 gives the fallback at cancellation) and can be trained by gradient descent, whereas the global rule selects one candidate through an argmax, so its output changes discontinuously with θ and has zero gradient almost everywhere; the selector is therefore trained with PhaseFuse-PS, and PhaseFuse-GH is used only at inference. We denote the complete fusion, from the candidates to the reconstruction, by $\widehat{\boldsymbol{x}} = G_\theta(\{\boldsymbol{v}_e\}_{e\in\mathcal{E}}, \boldsymbol{x}_0; \boldsymbol{y}, \boldsymbol{A})$, where $\boldsymbol{x}_0$ provides the reference and $\boldsymbol{y}, \boldsymbol{A}$ enter through the residuals.

3.4 SELF-SUPERVISED TRAINING

For each procedural object $\boldsymbol{x}_r^\star$, we simulate two observations $\boldsymbol{y}_r^{(a)}$ and $\boldsymbol{y}_r^{(b)}$ of $|\boldsymbol{A}\boldsymbol{x}_r^\star|^2$ with independent zero-mean noise. Only $\boldsymbol{y}_r^{(a)}$ enters the experts and the selector, producing $\widehat{\boldsymbol{x}}_{\theta,r} = G_\theta(\{\boldsymbol{v}_{e,r}\}, \boldsymbol{x}_{0,r}; \boldsymbol{y}_r^{(a)}, \boldsymbol{A})$, and $\boldsymbol{y}_r^{(b)}$ serves as the target:

$$\mathcal{L}_{\text{meas}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{r \in \mathcal{B}} \frac{\big\| |\boldsymbol{A}\widehat{\boldsymbol{x}}_{\theta,r}|^2 - \boldsymbol{y}_r^{(b)} \big\|_2^2}{\|\boldsymbol{y}_r^{(a)}\|_2^2}. \tag{6}$$

Since $\widehat{\boldsymbol{x}}_{\theta,r}$ and the normalization depend only on $\boldsymbol{y}_r^{(a)}$, the expectation of the numerator over the noise of $\boldsymbol{y}_r^{(b)}$ equals $\||\boldsymbol{A}\widehat{\boldsymbol{x}}_{\theta,r}|^2 - |\boldsymbol{A}\boldsymbol{x}_r^\star|^2\|_2^2$ plus the target-noise variance, which does not depend on θ. Minimizing Eq. (6) is therefore equivalent in expectation to minimizing the normalized error against the noiseless measurements, without access to $\boldsymbol{x}_r^\star$; this is the Noise2Noise principle (Lehtinen et al., 2018) applied in the intensity domain. This equivalence concerns expected risk and does not imply that every object is recovered. The full objective is $\mathcal{L} = \mathcal{L}_{\text{meas}} + \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\text{TV}}\mathcal{L}_{\text{TV}}$, where $\mathcal{L}_{\text{TV}}$ is the total variation of the reconstruction on its unit phasors and $\mathcal{L}_{\text{rot}}$ asks the fusion of quarter-turn-rotated candidates and reference to equal the rotated reconstruction up to a global phase (Appendix A.4). Only θ is optimized, the ground-truth objects are used only by the simulator, and training uses the single-pass fusion, in which each expert runs its full budget of N updates from $\boldsymbol{x}_0$.

3.5 RECURSIVE EXPERT FUSION

Recursive expert fusion splits the budget of N updates per expert into D decisions of $\tau = N/D$ updates and restarts the experts from the fused estimate after each decision,

$$\boldsymbol{x}_{d+1} = G_\theta\Big(\big\{\mathcal{F}_e^{(\tau)}(\boldsymbol{y}, \boldsymbol{A}; \boldsymbol{x}_d)\big\}_{e \in \mathcal{E}}, \boldsymbol{x}_d; \boldsymbol{y}, \boldsymbol{A}\Big), \qquad d = 0, \dots, D-1, \tag{7}$$

with output $\widehat{\boldsymbol{x}} = \boldsymbol{x}_D$. At each decision, $\boldsymbol{x}_d$ is both the starting point of the experts and the alignment reference, and $D = 1$ recovers the single pass. Since $D\tau = N$ for every D, recursion adds no expert iterations; the spectral expert lets the selector keep the current estimate at any decision; and experts with internal variables, such as the relaxation schedule of RAAR and the dual variables of LAD–ADMM, reinitialize them at each restart. Recursion changes inference only: the selector trained with the single pass is reused unchanged, and the added cost is D fusion steps instead of one, each of which aligns the candidates, computes their residuals, and evaluates the selector. Under PhaseFuse-GH, the recursion becomes a measurement-driven sequence of whole-image expert choices. If the selector places all weight on a single expert e at every decision, Eq. (7) reduces, up to a global phase, to that expert restarted from its own projected output,

$$\boldsymbol{z}_{d+1} = \mathcal{P}_{\mathbb{S}^1}\big(\mathcal{F}_e^{(\tau)}(\boldsymbol{y}, \boldsymbol{A}; \boldsymbol{z}_d)\big), \qquad \boldsymbol{z}_0 = \boldsymbol{x}_0, \tag{8}$$

which is how PhaseFuse behaves at $L = 3, 4$ (Section 4.4). Appendix C.2 also uses expert-specific block lengths τ_e, with $N_e = D\tau_e$. The selector can optionally be fitted to the recursive schedule by minimizing the sum over decisions of the loss of Eq. (6) evaluated at $\boldsymbol{x}_{d+1}$, with the state $\boldsymbol{x}_d$ detached between decisions and the expert outputs treated as constants.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and forward model. Images are encoded as phase-only objects: an intensity image $I \in [0,1]^{H \times W}$ becomes $\boldsymbol{x} = \exp\{j(2\pi\,\mathrm{vec}(I) - \pi)\}$, so the task is to recover the phase that encodes the image. Measurements follow Eq. (1) with L random CDP masks and additive Gaussian noise at a prescribed SNR (Appendix B). Unless stated otherwise, we use $L = 2$, 30 dB, and 128×128 images. We evaluate on all 200 images of the BSDS500 test split (Martin et al., 2001; Arbeláez et al., 2011), resized with bicubic interpolation, and, for cross-domain transfer, on 200 FFHQ validation images (Karras et al., 2019), resized in the same way. RGB images are reconstructed channel by channel with the grayscale selector, so they use three times as many measurements as grayscale images.

Table 1: Mean phase-PSNR (dB) on the 200 BSDS500 test images ($L = 2$, 30 dB, 128×128 unless stated). Experts use 200 updates from the shared initialization; Spectral is the expert that returns its input unchanged, here the shared spectral initialization, and ADMM is LAD–ADMM. The single pass uses $D = 1$, $\tau = 200$ and the recursive columns $D = 25$, $\tau = 8$, with the same selector. Best per row in bold; gains in the text are computed from unrounded means.

	Experts						Single pass	Recursive	
Condition	Spectral	PAP	RAAR	TWF	RAF	ADMM	PhaseFuse-PS	PhaseFuse-PS	PhaseFuse-GH
20 dB	18.31	21.16	12.89	23.07	4.87	11.70	25.16	**27.18**	26.62
30 dB	18.35	26.43	20.47	26.65	4.80	16.88	28.70	35.97	**36.45**
40 dB	18.35	28.59	25.48	27.51	4.79	20.64	30.36	44.09	**46.36**
64×64	17.31	25.01	20.05	25.06	4.81	16.13	27.04	**35.02**	33.84
256×256	19.19	27.30	20.62	27.76	4.80	17.30	29.85	35.17	**36.47**
RGB	18.31	26.41	20.42	26.65	4.80	16.88	28.66	35.97	**36.45**

Training. We train one selector per snapshot count $L \in \{1, 2, 3, 4\}$ on 10,000 procedural 128×128 images, each from a distinct IFS category (Section 3.1), with training SNR drawn uniformly in decibels from 15 to 45 dB and never provided to the experts or the selector. Each object keeps one SNR and one pair of noise realizations for the whole run, and its expert candidates are computed once and cached. All experiments use one fixed set of L masks (seed 0), shared by training and test. We use Adam with learning rate 3×10^{-4}, no weight decay, batch size 128, gradient clipping at 1.0, $\lambda_{\mathrm{rot}} = 0.1$, $\lambda_{\mathrm{TV}} = 10^{-3}$, and $T = 0.6$, and train for 140 epochs, which takes 1.2 h on one RTX 3090. The number of epochs was fixed before evaluation; fifty disjoint procedural categories monitor the measurement loss, and no checkpoint is selected on a test metric. The resolution study trains separate selectors at 64×64 and 256×256 with the same protocol, and the temperature study trains separate selectors at $T = 0.3$ and $T = 1.0$. Architecture and generator details are given in Appendix A.

Experts and inference. The five iterative experts receive $N = 200$ updates each from the shared filtered spectral initialization, computed with 50 power iterations (Appendix A.2); Appendix C.6 shows why a shared initialization is used. The single-pass fusion uses $D = 1$, $\tau = 200$, and recursive fusion uses $D = 25$, $\tau = 8$, with the same selector and no retraining. Within each condition, every method receives the same image, masks, noisy measurements, and initialization.

Metric. We report mean phase-PSNR over images, computed after global phase alignment with peak phase error π (Appendix B); for RGB, channel scores are averaged. The best expert of a condition is the strongest individual expert in that condition, identified after all reconstructions are fixed.

4.2 Main results

Table 1 compares PhaseFuse with every expert across noise levels, resolutions, and color representations. Under the single pass, PhaseFuse-PS exceeds the best expert by 1.8 to 2.1 dB in every condition, and by 2.05 dB at the reference setting. Recursive fusion, with the same budget of 200 updates per expert and the same selector, raises this margin at the reference setting to 9.32 dB for PhaseFuse-PS and 9.80 dB for PhaseFuse-GH; across conditions, the better of the two rules gains between 4.1 dB at 20 dB SNR and 17.8 dB at 40 dB SNR.

The improvement holds image by image: every test image improves over the single pass at every recursive schedule ($D > 1$) in Table 2, by at least 1.40 dB and by at least 2.49 dB at $D = 25$, and the standard deviation across images falls from 2.62 to 0.23 dB, so recursion makes the reconstruction more predictable as well as more accurate. Figure 2 shows typical cases. Independent RGB reconstruction matches grayscale, and in every condition the selector suppresses RAF, which fails under the shared initialization and fixed step size. The added cost is 25 fusion steps instead of one, which increases inference time from 34.4 to 60.7 ms per image ($1.76\times$, batch 50, one RTX 3090), while the analytic budget is unchanged.

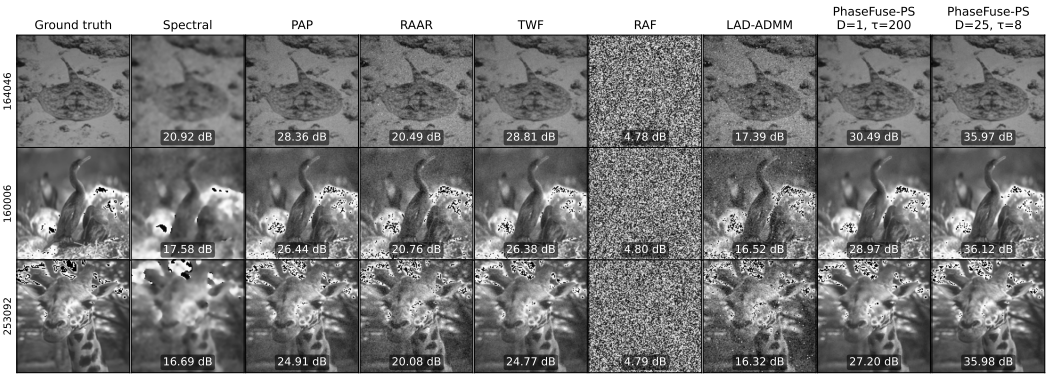


Figure 2: Single-pass against recursive PhaseFuse-PS ($L = 2$, 30 dB, 128×128). Rows are BSDS500 test images at the 25th, 50th, and 75th percentiles of the per-image gain of recursive over single-pass fusion. All columns of a row share $\boldsymbol{y}$, $\boldsymbol{A}$, and $\boldsymbol{x}_0$; each expert uses 200 updates in one run, both fusion columns spend 200 updates per iterative expert, and the Spectral column is the shared initialization. Labels are phase-PSNR after global phase alignment.

Table 2: Fusion schedule for PhaseFuse-PS at a fixed budget of $D\tau = 200$ updates per expert, with one single-pass-trained selector (200 BSDS500 test images, $L = 2$, 30 dB). σ is the standard deviation across images, and Min. gain is the smallest per-image gain over the single pass; every image improves at every $D > 1$.

D	1	2	4	5	8	10	20	25	40	50	100	200
τ	200	100	50	40	25	20	10	8	5	4	2	1
dB	28.70	32.20	33.91	34.27	34.86	35.11	35.80	35.97	36.28	36.30	**36.34**	36.13
σ	2.62	1.64	0.87	0.71	0.47	0.39	0.26	**0.23**	0.38	0.42	0.46	0.54
Min. gain	—	1.40	1.70	1.75	1.89	2.00	2.41	2.49	2.82	2.83	2.73	2.40

4.3 Effect of the fusion schedule and analytic budget

Table 2 varies the number of decisions at a fixed total budget. Accuracy rises steeply up to $D \approx 25$, remains flat up to $D = 100$, and drops at $D = 200$, where each expert performs a single update per block. Since the selector is the same in every column and the analytic budget is fixed, the gain over the single pass comes from the schedule alone.

The selector can also be fitted to the schedule it runs at by minimizing Eq. (6) through the fusion at every decision. Starting from the single-pass checkpoint and training for 20 epochs at $D = 25$, $\tau = 8$ (batch 256, learning rate 3×10^{-4}) raises $D = 25$ from 35.97 to 37.25 dB and $D = 100$ from 36.34 to 37.61 dB, at the cost of 2.39 dB at $D = 1$, which shows that the refitted selector specializes to the recursive schedule. All other results use the single-pass selectors.

Changing the analytic budget at test time confirms that more expert work is not the source of the gain (Appendix C.2). Raising the budget to 500 updates per expert lowers PhaseFuse-PS to 35.06 dB, and every expert except RAF also scores lower at 500 updates than at 200. Letting each expert stop where its measurement error on procedural objects is lowest reaches 36.03 dB with about half the updates, although part of this gain reflects a weaker best-expert baseline, since calibration also lowers TWF from 26.65 to 25.63 dB.

4.4 Per-pixel versus image-level routing

Neither routing rule dominates under recursion (Table 1). PhaseFuse-GH is ahead by 0.49 dB at 30 dB, 2.27 dB at 40 dB, and 1.30 dB at 256×256, while PhaseFuse-PS is ahead by 0.56 dB at 20 dB and 1.17 dB at 64×64. The learned weights across snapshot counts are consistent with this pattern (Table 3). With one snapshot, the per-pixel weights spread almost evenly over the spectral, PAP, and TWF experts; with two, they concentrate on TWF; and with three or four, TWF is the

Table 3: Mean expert weights at the final decision of recursive fusion ($D = 25$, $\tau = 8$, 30 dB, 128×128) over all pixels of the 200 test images, with the selector trained at each L. Entropy is per pixel, normalized by $\log 6$, and averaged. For PhaseFuse-GH the weights are one-hot, so the columns are the fraction of images selecting each expert. Values are rounded; at $L = 3, 4$ the TWF weight exceeds 0.99999 and the entropy is below 2×10^{-5} at every decision.

Rule	L	Entropy	Spectral	PAP	RAAR	TWF	RAF	ADMM
PhaseFuse-PS	1	0.61	0.30	0.31	0.07	0.30	0.02	0.00
PhaseFuse-PS	2	0.48	0.21	0.25	0.01	0.52	0.00	0.01
PhaseFuse-GH	1	0.00	0.12	0.40	0.00	0.49	0.00	0.00
PhaseFuse-GH	2	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Both	3, 4	0.00	0.00	0.00	0.00	1.00	0.00	0.00

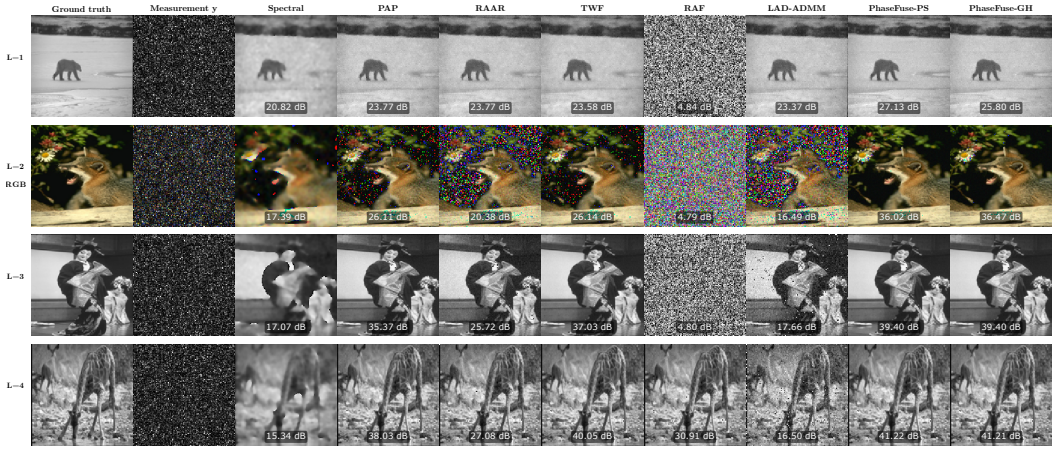


Figure 3: BSDS500 examples at 128×128 and 30 dB, one row per snapshot count, with the $L = 2$ example in RGB and the others in grayscale. Columns show the ground truth, the first coded diffraction pattern of $\boldsymbol{y}$ (each channel scaled to its 99.5th percentile, for display only), the shared initialization (Spectral), the experts after 200 updates, and both rules under recursive fusion ($D = 25$, $\tau = 8$). At $L = 3, 4$ both rules place nearly all weight on TWF and agree to within 0.01 dB. Labels are phase-PSNR; per-panel routing is in Table 16.

dominant expert at every pixel, decision, and image and holds at least 0.99999 of the mean weight, so the two rules produce nearly identical reconstructions (Figure 3) and PhaseFuse reduces to TWF restarted from its projected output, Eq. (8), which still gains 2.44 and 1.11 dB over TWF at $L = 3$ and 4 (Appendix C.1). This choice is not trivial: without ground truth or labeled data the best expert cannot be established directly, and it changes with the condition (PAP at $L = 1$ and at 40 dB, TWF elsewhere), so the selector must identify it from the measurements alone. We hypothesize that mixing responds to ambiguity in the measurements: when several experts are plausible, hedging across them pays, and when one expert is clearly right, hedging only dilutes it; testing this at a controlled measurement ratio and noise level remains open.

Table 4 isolates what the learned weights contribute. Uniform weights fall 4.59 dB below TWF alone, because the failed RAF expert still receives weight $1/6$ at every pixel, and random routing collapses. Removing RAF by hand or averaging only PAP and TWF recovers most of the gain, but we chose these subsets after inspecting the test results, so they bound what a fixed subset can reach rather than providing a deployable baseline; the selector identifies the failing experts from the measurements alone. The case for per-pixel routing specifically depends on the schedule: under the single pass, the per-pixel blend beats a separately trained selector that produces one weight vector per image by 0.77 dB (28.70 against 27.93 dB), whereas under recursion, averaging the PhaseFuse-PS logits over the image before the softmax is slightly ahead of the per-pixel blend (36.09 against 35.97 dB), suggesting that repeating the decision substitutes for spatial freedom within a single decision. Selectors trained at $T = 0.3$ and $T = 1.0$ reach 35.50 and 35.72 dB, against 35.97 dB at $T = 0.6$, so the result is robust to the softmax temperature.

Table 4: Routing rules under recursive fusion ($D = 25$, $\tau = 8$, $L = 2$, 30 dB, 200 images). Δ is the difference to TWF (26.65 dB). All rules except uniform and random read the logits of the PhaseFuse-PS selector; the image-level blend averages them over the image before the softmax. Random rules draw a new expert at every decision and average three draws. All rules share the same alignment and projection. Oracle rules (uniform without RAF; mean of PAP and TWF) were chosen after inspecting the test results.

Rule	dB	Δ	Rule	dB	Δ
PhaseFuse-PS (per-pixel blend)	35.97	$+9.32$	Uniform	22.06	-4.59
PhaseFuse-GH (image-level top-1)	**36.45**	$+9.80$	Random, per pixel	8.21	-18.44
Image-level blend	36.09	$+9.44$	Random, per image	5.13	-21.52
Per-pixel top-1	34.20	$+7.55$	Oracle: no RAF	33.17	$+6.52$
			Oracle: PAP and TWF	35.41	$+8.76$

Blending versus sampling. Reading the logits as a per-pixel categorical policy and fine-tuning the selector with REINFORCE, with the measurement objective on the second noise realization as reward, makes it worse than its warm start under every schedule (Table 15, Appendix C.8).

4.5 Training data and training length

To measure what training on procedural objects costs, we trained a second selector with an identical protocol on the 200 BSDS500 training images, cycled 50 times to fill the same 10,000 training rows, each with its own noise realization, so that the number of optimizer steps matches while the number of distinct objects is 50 times smaller (Table 13 in Appendix C.7). Under the single pass, the natural-image selector is ahead on both test sets, 29.49 against 28.70 dB on BSDS500 and 31.87 against 31.00 dB on FFHQ. Under recursive fusion, the procedural selector leads on BSDS500, 35.97 against 35.48 dB, and on FFHQ, which neither selector has seen, 36.15 against 35.45 dB. The differences are 0.49 and 0.70 dB and both selectors are far above every expert, so in these experiments the procedural restriction does not reduce accuracy once the selector is applied recursively.

Training length behaves similarly (Table 14): single-pass accuracy varies little across checkpoints, between 28.61 and 28.83 dB, whereas under recursion the epoch-20 checkpoint, obtained after 0.17 h, already reaches 36.11 dB against 35.97 dB for the predeclared epoch-140 checkpoint, which we report everywhere else.

5 Limitations and conclusion

Our comparison is restricted to analytic solvers under a shared initialization and budget; we do not compare with learned PR methods, including untrained and measurement-supervised networks that avoid clean targets. With three or four snapshots the selector places nearly all weight on TWF, so the method reduces to a schedule for a single expert, chosen by the selector without labels: its single-pass gain over TWF vanishes, and its recursive gain of 2.44 and 1.11 dB comes from restarting TWF rather than from mixing (Appendix C.1). Per-pixel routing also peaks at 128×128 under recursion. All experiments use simulated CDP measurements of phase-only objects with Gaussian noise, one fixed set of masks shared by training and test, and one training seed and noise realization per condition; general complex-valued objects, other noise models and operators, and experimental data remain open.

We presented PhaseFuse, a self-supervised selector that combines phase-aligned analytic reconstructions using procedural training objects. In simulated phase-only CDP, repeated fusion raises the margin over the strongest configured expert from 2.05 to 9.32 dB at two snapshots and 30 dB, improves every test image, and preserves each expert's update count, at 1.76 times the inference time of single-pass fusion. Spatial blending and image-level selection have different strengths across conditions, and routing concentrates as the snapshot count grows. Because the recursive gain persists at three and four snapshots with nearly all weight on one expert, these results support adaptive solver schedules but do not yet isolate learned fusion from repeated unit-modulus projection and solver restarts; tuned, compute-matched baselines and scientific and experimental data remain necessary.

AI USE STATEMENT

We used generative AI tools for three purposes. First, to aid and polish the writing, including the translation of the authors' drafts from Spanish. Second, to draft sections of the paper from the authors' notes and results. Third, in research ideation and execution: an AI coding assistant helped propose ablations, implement and run the experiments, generate tables and figures, and check the reported numbers against the saved experiment outputs. We have reviewed all AI-assisted text, code, and results, and we take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a self-supervised method for phase retrieval from simulated coded diffraction measurements and does not involve human subjects, user studies, or the collection of new data. The selector is trained only on procedurally generated images, which carry no personal or copyrighted content. For evaluation, we use the public BSDS500 test split and FFHQ validation images, the latter released by NVIDIA under CC BY-NC-SA 4.0 with individual images under Creative Commons BY 2.0, BY-NC 2.0, Public Domain Mark 1.0, CC0 1.0, or U.S. Government Works licenses, and we use them for non-commercial research only. FFHQ contains photographs of human faces; we use them only as sources of phase-only test objects, report aggregate reconstruction metrics, and perform no identification, attribute prediction, or demographic analysis. Because PhaseFuse restricts its output to projected blends of analytic candidates rather than generating content from a learned image prior, it is less prone to inventing image structure, which is desirable in scientific imaging, where reconstructions may inform downstream conclusions; this restriction does not, however, guarantee the absence of reconstruction artifacts. We do not foresee direct harmful applications beyond those of computational imaging in general. Nevertheless, reconstructions from real instruments should be validated against physical measurements, since our experiments are limited to simulated data (Section 5). Training and evaluation require a single consumer GPU, which keeps the computational cost of the study modest.

REPRODUCIBILITY STATEMENT

The forward model and the problem setting are defined in Section 2, and every component of the method, including the procedural generator, the expert library, the global phase alignment, the selector, the training objective, and the recursive schedule, is specified in Section 3. Section 4.1 lists the datasets, splits, noise levels, optimizer, hyperparameters, number of epochs, hardware, and training time. Appendix A provides the generator parameters and random seeds, the exact update rule and constants of every expert and of the shared initialization, the selector architecture, and the definitions of the regularizers, and Appendix B defines the evaluation metric, the noise model, and the image preprocessing. All checkpoints and the number of training epochs were fixed before evaluation, and Appendix C reports the complete results of every condition, including standard deviations, the procedural monitor, and the hyperparameter study.

REFERENCES

Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. doi: 10.1109/TPAMI.2010.161.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015. doi: 10.1016/j.acha.2014.09.004.

Chien-Chun Chen, Jianwei Miao, C. W. Wang, and T. K. Lee. Application of optimization technique to noncrystalline x-ray diffraction microscopy: Guided hybrid input-output method. *Physical Review B*, 76:064113, 2007. doi: 10.1103/PhysRevB.76.064113.

Pengwen Chen, Albert Fannjiang, and Gi-Ren Liu. Phase retrieval with one or two diffraction patterns by alternating projections with the null initialization. *Journal of Fourier Analysis and Applications*, 24:719–758, 2018. doi: 10.1007/s00041-017-9536-8.

Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5): 822–883, 2017. doi: 10.1002/cpa.21638.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Alessandro Colombo, Davide Emilio Galli, Liberato De Caro, Francesco Scattarella, and Elvio Carlino. Facing the phase problem in coherent diffractive imaging via memetic algorithms. *Scientific Reports*, 7:42236, 2017.

Alessandro Colombo et al. SPRING, an effective and reliable framework for image reconstruction in single-particle coherent diffraction imaging. *npj Computational Materials*, 11:265, 2025. doi: 10.1038/s41524-025-01661-y.

David K. Hammond and Eero P. Simoncelli. A machine learning framework for adaptive combination of signal denoising methods. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 6, pp. 29–32, 2007.

O. Hoidn, A. A. Mishra, and A. Mehta. Physics constrained unsupervised deep learning for rapid, high resolution scanning coherent diffraction reconstruction. *Scientific Reports*, 13:22789, 2023. doi: 10.1038/s41598-023-48351-7.

Andrés Jerez, Juan Estupiñán, Jorge Bacca, and Henry Arguello. Deep unrolled single snapshot phase retrieval via non-convex formulation and phase mask design. *IEEE Journal of Selected Topics in Signal Processing*, 18(4):694–703, 2024. doi: 10.1109/JSTSP.2024.3395979.

Xue Jiang, Hing Cheung So, and Xingzhao Liu. Robust phase retrieval via ADMM with outliers. *arXiv preprint arXiv:1702.06157*, 2017. URL https://arxiv.org/abs/1702.06157.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Mehmet Onurcan Kaya and Figen S. Oktem. I2I-PR: Deep iterative refinement for phase retrieval using image-to-image diffusion models. *Digital Signal Processing*, 180:106184, 2026. doi: 10.1016/j.dsp.2026.106184.

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 *PMLR*, pp. 2965–2974, 2018.

D. Russell Luke. Relaxed averaged alternating reflections for diffraction imaging. *Inverse Problems*, 21(1):37–50, 2005. doi: 10.1088/0266-5611/21/1/004.

D. Russell Luke. ProxToolbox, version 1.0.3. Working Group in Variational Analysis, University of Göttingen, 2015. URL https://num.math.uni-goettingen.de/proxtoolbox/versions/1.0.3/.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV)*, volume 2, pp. 416–423, 2001. doi: 10.1109/ICCV.2001.937655.

Christopher Metzler, Phillip Schniter, Ashok Veeraraghavan, and Richard Baraniuk. prDeep: Robust phase retrieval with a flexible deep network. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3501–3510. PMLR, 2018.

Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342–344, 1999. doi: 10.1038/22498.

Rick P. Millane. Phase retrieval in crystallography and optics. *Journal of the Optical Society of America A*, 7(3):394–411, 1990. doi: 10.1364/JOSAA.7.000394.

Ryo Nakamura, Hirokatsu Kataoka, Sora Takashima, Edgar Josafat Martinez Noriega, Rio Yokota, and Nakamasa Inoue. Pre-training vision transformers with very limited synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20360–20369, 2023. URL https://ryoo-nakamura.github.io/One-instance-FractalDB/.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*

(MICCAI), volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. doi: 10.1109/MSP.2014.2352673.

Ayan Sinha, Justin Lee, Shuai Li, and George Barbastathis. Lensless computational imaging through deep learning. *Optica*, 4(9):1117–1125, 2017. doi: 10.1364/OPTICA.4.001117.

Pierre Thibault, Martin Dierolf, Andreas Menzel, Oliver Bunk, Christian David, and Franz Pfeiffer. High-resolution scanning X-ray diffraction microscopy. *Science*, 321(5887):379–382, 2008. doi: 10.1126/science.1158573.

Fei Wang, Yaoming Bian, Haichao Wang, Meng Lyu, Giancarlo Pedrini, Wolfgang Osten, George Barbastathis, and Guohai Situ. Phase imaging with an untrained neural network. *Light: Science & Applications*, 9:77, 2020. doi: 10.1038/s41377-020-0302-3.

Gang Wang, Georgios B. Giannakis, Yousef Saad, and Jie Chen. Phase retrieval via reweighted amplitude flow. *IEEE Transactions on Signal Processing*, 66(11):2818–2833, 2018. doi: 10.1109/TSP.2018.2818077.

Y. Yao, H. Chan, S. Sankaranarayanan, P. Balaprakash, R. J. Harder, and M. J. Cherukara. AutoPhaseNN: Unsupervised physics-aware deep learning of 3D nanoscale Bragg coherent diffraction imaging. *npj Computational Materials*, 8:124, 2022. doi: 10.1038/s41524-022-00803-w.

A IMPLEMENTATION DETAILS

A.1 PROCEDURAL GENERATOR

The generator follows the fixed-formula, one-representative-per-category idea of 2D-OFDB (Nakamura et al., 2023), whose categories were introduced for visual pre-training; here they only index generated phase objects, and neither a category classifier nor pretrained weights are used. Each category samples $J \in \{2, \dots, 7\}$ affine maps with coefficients in $[-1, 1]$. For square side S, each candidate orbit contains $N_q = \max(12000, 3S^2)$ points, and the first $\max(600, \lfloor N_q/10 \rfloor)$ are discarded. Systems with nonfinite coordinates, magnitude above 10^{100}, or degenerate extent are rejected. Up to 64 candidate systems are searched per category, and the first with occupancy of at least 0.20 on a 64×64 grid is accepted; otherwise, the finite system with the highest occupancy is kept, so 20% is a search target rather than a guaranteed fill rate. Random affine coefficients alone do not guarantee contractivity.

Each occupied pixel of the representative image is replaced by a nonempty binary 3×3 patch, specified by an independent integer in $\{1, \dots, 511\}$; overlapping contributions are summed, and $\log(1 + \text{count})$ followed by min–max normalization gives grayscale intensities. Two independently seeded augmentations enter Eq. (3) with $\alpha \sim \mathrm{Unif}[0.15, 0.45]$; the coarse image is area-resampled to $\max(2, \lfloor S/4 \rfloor)$ and bicubically restored to S, and a deterministic category-dependent flip is applied. One composite image is stored per category. The IFS search is seeded by $81000 +$ category, pattern augmentation by $91000000 +$ index $+ 1000003v$ for variants $v = 0, 1$, and the mixing weight by $92000 +$ index. Training categories start at zero, and the 50 monitoring categories start at 1,000,000. The renderer, point budget, and multiscale composition are local adaptations and do not reproduce the original OFDB images.

A.2 INITIALIZATION AND EXPERT UPDATES

We give the implemented single-channel maps. Let $\boldsymbol{b} = \sqrt{\max(\boldsymbol{y}, 0)}$ and $\mathrm{ph}(z) = z/\max(|z|, \epsilon)$ with $\epsilon = 10^{-8}$. The initializer keeps the largest half of the nonnegative intensities in a binary support $\boldsymbol{s}$ and repeats

$$\boldsymbol{q}^{j+1} = \frac{\mathcal{G}\{\boldsymbol{A}^H[\boldsymbol{s} \odot (\boldsymbol{A}\boldsymbol{q}^j)]\}}{\max(\|\mathcal{G}\{\boldsymbol{A}^H[\boldsymbol{s} \odot (\boldsymbol{A}\boldsymbol{q}^j)]\}\|_2, \epsilon)}, \qquad j = 0, \dots, 49, \tag{9}$$

where $\mathcal{G}$ is a fixed 5×5 Gaussian low-pass filter with standard deviation 1.25. The seeded complex initial vector is normalized, and the final vector is scaled by $\sqrt{\max(\sum_i \max(y_i, 0), \epsilon)}$, the signal norm implied by $\boldsymbol{A}^H\boldsymbol{A} = \boldsymbol{I}_n$.

PAP. The object-domain map is $\boldsymbol{x}^{k+1} = \boldsymbol{A}^H[\boldsymbol{b} \odot \mathrm{ph}(\boldsymbol{A}\boldsymbol{x}^k)]$, extended in parallel over the L masks; the shared initialization replaces the null initializer of Chen et al. (2018).

RAAR. With $P_M(\boldsymbol{u}) = \boldsymbol{b} \odot \mathrm{ph}(\boldsymbol{u})$, $R_M = 2P_M - I$, and $P_R = \boldsymbol{A}\boldsymbol{A}^H$, the measurement-space update follows the reflection ordering of the ProxToolbox release (version 1.0.3) (Luke, 2015) of the method of Luke (2005),

$$\boldsymbol{u}^{k+1} = \tfrac{1}{2}\big[\beta_k\{2P_R R_M(\boldsymbol{u}^k) - R_M(\boldsymbol{u}^k)\} + (1-\beta_k)R_M(\boldsymbol{u}^k) + \boldsymbol{u}^k\big], \qquad \boldsymbol{x}^{k+1} = \boldsymbol{A}^H\boldsymbol{u}^{k+1}, \tag{10}$$

with $\boldsymbol{u}^0 = \boldsymbol{A}\boldsymbol{x}_0$ and $\beta_k = 0.75 + (0.99999 - 0.75)\{1 - e^{-(k/7)^3}\}$.

TWF. Let $\boldsymbol{u}^k = \boldsymbol{A}\boldsymbol{x}^k$, $\boldsymbol{h}^k = |\boldsymbol{u}^k|^2 - \max(\boldsymbol{y}, 0)$, and $\bar{h}^k = m^{-1}\sum_i |h_i^k|$. A measurement is kept when $0.3 \le \sqrt{m}|u_i^k|/\max(\|\boldsymbol{x}^k\|_2, \epsilon) \le 5$ and $|h_i^k| \le 5\sqrt{m}\,\bar{h}^k|u_i^k|/\max(\|\boldsymbol{x}^k\|_2, \epsilon)$. With indicator vector $\boldsymbol{\chi}^k$, the update is

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - 0.2\,\boldsymbol{A}^H\Big[\boldsymbol{\chi}^k \odot \frac{2\boldsymbol{h}^k \odot \boldsymbol{u}^k}{\max(|\boldsymbol{u}^k|^2, \epsilon)}\Big]. \tag{11}$$

The truncation and Poisson-gradient form follow Chen & Candès (2017); the native truncated spectral initialization is replaced by the shared one.

RAF. With $r_i^k = |u_i^k| / \max(b_i, \epsilon)$ and $w_i^k = r_i^k/(r_i^k+10)$, the update is $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - 5\,\boldsymbol{A}^H[\boldsymbol{w}^k \odot \{\boldsymbol{u}^k - \boldsymbol{b} \odot \mathrm{ph}(\boldsymbol{u}^k)\}]$, following the authors' release (Wang et al., 2018). The fixed step can be unstable under the shared initialization, and its poor scores are retained. TWF and RAF keep the previous iterate if an update contains nonfinite entries.

LAD–ADMM. With residual variable $\boldsymbol{z}$, unscaled dual $\boldsymbol{\lambda}$, and $\rho = 1$,

$$\boldsymbol{x}^{k+1} = \boldsymbol{A}^H\big[(\boldsymbol{b} + \boldsymbol{z}^k - \boldsymbol{\lambda}^k/\rho) \odot \mathrm{ph}(\boldsymbol{A}\boldsymbol{x}^k)\big], \tag{12}$$

$$\boldsymbol{z}^{k+1} = \mathrm{soft}_{1/\rho}\big(|\boldsymbol{A}\boldsymbol{x}^{k+1}| - \boldsymbol{b} + \boldsymbol{\lambda}^k/\rho\big), \tag{13}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \rho\big(|\boldsymbol{A}\boldsymbol{x}^{k+1}| - \boldsymbol{b} - \boldsymbol{z}^{k+1}\big), \qquad \boldsymbol{z}^0 = \boldsymbol{\lambda}^0 = \boldsymbol{0}, \tag{14}$$

which uses one alternating phase and least-squares step followed by the LAD soft threshold and dual update of Jiang et al. (2017). All updates are written under the common normalized operator; we do not claim that the original papers used the same FFT scaling, initialization, or sensing conditions.

A.3 Selector architecture and global-phase symmetry

Table 5: Selector architecture. Hidden convolutions use 3×3 kernels and ReLU; the output head is linear.

Stage	Channels	Spatial size
Input features	22	$H \times W$
Encoder 1: two convolutions	$22 \to 48 \to 48$	$H \times W$
Max-pool, encoder 2: two convolutions	$48 \to 96 \to 96$	$H/2 \times W/2$
Max-pool, bottleneck convolution	$96 \to 96$	$H/4 \times W/4$
Upsample, concatenate encoder 2, convolution	$192 \to 96$	$H/2 \times W/2$
Upsample, concatenate encoder 1, convolution	$144 \to 48$	$H \times W$
Logit head	$48 \to 6$	$H \times W$

The 22 input channels are the real and imaginary parts of the $K = 6$ aligned candidates relative to the reference ($2K$), the real and imaginary parts of the horizontal and vertical neighbor phasors $r(p+e)\overline{r(p)}$ of the reference, padded with 1 at the last column and row (4), and the K broadcast residuals. Upsampling uses nearest neighbors, and spatial sizes assume sides divisible by four, as in all experiments. The weights are shared across pixels, batch elements, and color channels, and one model is trained per snapshot count.

To verify the symmetry, rotate the candidates independently, $\boldsymbol{u}'_e = e^{j\beta_e}\boldsymbol{u}_e$, and the reference, $\boldsymbol{r}' = e^{j\alpha}\boldsymbol{r}$. If $\langle \boldsymbol{r}, \boldsymbol{u}_e\rangle \neq 0$, then $\arg\langle \boldsymbol{r}', \boldsymbol{u}'_e\rangle = \beta_e - \alpha + \arg\langle \boldsymbol{r}, \boldsymbol{u}_e\rangle$ and $\widetilde{\boldsymbol{v}}'_e = e^{j\alpha}\widetilde{\boldsymbol{v}}_e$, so $\widetilde{v}'_e(p)\overline{r'(p)} = \widetilde{v}_e(p)\overline{r(p)}$. The reference phase differences and the residuals are also unchanged, so the network receives the same input and produces the same weights, and the blend satisfies $G_\theta(\{e^{j\beta_e}\boldsymbol{v}_e\}, e^{j\alpha}\boldsymbol{x}_0; \boldsymbol{y}, \boldsymbol{A}) = e^{j\alpha}G_\theta(\{\boldsymbol{v}_e\}, \boldsymbol{x}_0; \boldsymbol{y}, \boldsymbol{A})$. The selector is thus invariant and the reconstruction equivariant on nondegenerate inputs. In the implementation, near-zero candidate or reference entries are replaced by the unit phasor 1, a near-zero correlation uses the identity rotation, and a near-zero fused sum falls back to the normalized reference; exact invariance is not claimed for these degenerate cases.

A.4 Training regularizers and optimization

For a quarter-turn rotation $\mathcal{R}$, let $\widehat{\boldsymbol{x}}^{\mathcal{R}}_{\theta,r}$ be the output of G_θ on the rotated candidates and reference, with the residuals of the unrotated candidates. The rotational term is

$$\mathcal{L}_{\mathrm{rot}} = \frac{1}{|\mathcal{B}|}\sum_{r\in\mathcal{B}}\left[1 - \frac{|\langle \widehat{\boldsymbol{x}}^{\mathcal{R}}_{\theta,r}, \mathcal{R}\widehat{\boldsymbol{x}}_{\theta,r}\rangle|}{\max\{\|\widehat{\boldsymbol{x}}^{\mathcal{R}}_{\theta,r}\|_2 \|\mathcal{R}\widehat{\boldsymbol{x}}_{\theta,r}\|_2, \epsilon\}}\right], \tag{15}$$

and the implementation cycles through 90°, 180°, and 270° rotations across updates. The term regularizes the fusion map and does not assume that the mask operator commutes with rotation. The total variation term averages $|\widehat{x}(p) - \widehat{x}(q)|$ over horizontal neighbors and over vertical neighbors separately and sums the two means, comparing unit phasors directly so that no discontinuity appears

at $\pm\pi$. Gradients of the mean batch loss update the single shared selector, and the expert outputs are fixed inputs. The 256×256 selector accumulates weighted microbatch gradients to keep an effective batch of 128.

B Evaluation metrics and conventions

Phase-PSNR. We remove the unavoidable global phase of each channel before scoring, using the correlation-based alignment $\phi_c = \arg\langle \boldsymbol{x}_c^\star, \widehat{\boldsymbol{x}}_c \rangle$; the wrapped phase error and phase-PSNR are then

$$e_c(p) = \mathrm{wrap}_{[-\pi,\pi]}\left[\arg(e^{-j\phi_c}\widehat{x}_c(p)) - \arg(x_c^\star(p))\right], \tag{16}$$

$$\mathrm{PSNR}_{\phi,c} = 10\log_{10}\frac{\pi^2}{n^{-1}\sum_{p=1}^{n} e_c(p)^2}, \tag{17}$$

which uses peak phase error π and differs from intensity PSNR. The score of an image with C channels is $C^{-1}\sum_c \mathrm{PSNR}_{\phi,c}$. Dataset means and population standard deviations are computed over image scores, and gains are differences of paired means. Best-over-iteration values are retrospective diagnostics and never select an output, train the selector, or select a checkpoint.

Measurement SNR. With $\boldsymbol{s}_c = |\boldsymbol{A}\boldsymbol{x}_c^\star|^2$, the simulator draws $\boldsymbol{\omega}_c = \sigma_c \boldsymbol{\xi}_c$ with $\boldsymbol{\xi}_c \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$ and $\sigma_c^2 = m^{-1}\|\boldsymbol{s}_c\|_2^2 / 10^{\mathrm{SNR_{dB}}/10}$, which fixes the expected noise power. The noise level is not passed to any expert or to the selector. Negative noisy intensities are kept in the training loss, and experts that require amplitudes use $\sqrt{\max(\boldsymbol{y}_c, 0)}$.

Routing statistics. The dominant-expert share of expert e is the percentage of pixels whose largest weight is e, and the mean weight averages $w_e(p)$ directly; a 100% dominant share does not imply unit weights. The top weight is the mean over pixels of $\max_e w_e(p)$. Normalized entropy averages $-\sum_e w_e \log w_e / \log K$ over pixels, and the spatial spread is the mean within-image standard deviation of the expert weights.

Image preparation and display. Photographs are resized as whole images with bicubic interpolation before phase encoding, and RGB channels are processed independently with the same masks. For display only, reconstructions are mapped to $[0, 1]$ through $(\arg(e^{-j\phi_c}\widehat{x}_c) + \pi)/(2\pi)$, so wrapping near the endpoints can create visible differences even for small phase errors. Ground truth is accessed only after the reconstructions to be scored are fixed.

C Additional results

C.1 Single-pass results across snapshot counts

Under the single pass, PhaseFuse-PS improves over the best expert by 0.58 dB at $L = 1$ and 2.05 dB at $L = 2$, and ties TWF at $L = 3$ and $L = 4$, with unrounded differences of $+0.0002$ and -0.0003 dB (Table 6). RAF is unstable for $L \leq 3$ under the shared initialization and fixed step size, and the selector suppresses it. Under recursion (last two rows of Table 6), the better rule gains 3.12, 9.80, 2.44, and 1.11 dB over the best expert at $L = 1$ to 4. At $L = 3, 4$ nearly all weight goes to TWF, so this gain comes from restarting TWF from its projected output, Eq. (8), rather than from mixing.

Table 6: Single-pass fusion ($D = 1$, $\tau = 200$) across snapshot counts on the 200 BSDS500 test images (grayscale, 128×128, 30 dB): mean phase-PSNR $\pm$ population standard deviation. The Spectral row is the shared initialization, and three decimals are kept to show the numerical ties at $L = 3, 4$. Bold marks the best single-pass mean. The last two rows use recursive fusion ($D = 25$, $\tau = 8$) with the selector trained at each L and report means only.

Method	$L = 1$	$L = 2$	$L = 3$	$L = 4$
Spectral	17.676 ± 2.296	18.345 ± 2.405	18.581 ± 2.448	18.703 ± 2.468
PAP	20.597 ± 2.491	26.432 ± 2.210	35.444 ± 0.095	38.043 ± 0.079
RAAR	20.597 ± 2.491	20.467 ± 0.764	25.959 ± 0.169	27.408 ± 0.313
TWF	20.362 ± 2.432	26.650 ± 2.531	$\mathbf{37.004} \pm 0.149$	$\mathbf{40.172} \pm 0.083$
RAF	4.810 ± 0.016	4.798 ± 0.016	4.790 ± 0.013	35.382 ± 2.503
LAD–ADMM	20.339 ± 2.399	16.884 ± 1.222	17.677 ± 0.121	16.638 ± 0.079
PhaseFuse-PS	$\mathbf{21.174} \pm 2.729$	$\mathbf{28.696} \pm 2.620$	$\mathbf{37.004} \pm 0.149$	$\mathbf{40.172} \pm 0.083$
PhaseFuse-PS, recursive	23.714	35.965	39.440	41.278
PhaseFuse-GH, recursive	23.353	36.453	39.441	41.278

C.2 Analytic budget

Table 7 changes the analytic budget at test time with the same single-pass selector at $D = 25$. The calibrated budget lets each expert stop where its measurement error on procedural objects, evaluated against an independent noise realization, is lowest, giving 5, 3, 10, 1, and 2 updates per decision for PAP, RAAR, TWF, RAF, and LAD–ADMM. This generalizes Eq. (7) to expert-specific block lengths τ_e, for $N_e = D\tau_e = 125$, 75, 250, 25, and 50 updates, 525 in total against 1,000 for 200 each; the spectral expert performs no updates. Calibration rescues RAF, from 4.80 to 20.55 dB, and LAD–ADMM, from 16.88 to 24.83 dB, while TWF drops from 26.65 to 25.63 dB, so part of the larger margin reflects a weaker best-expert baseline rather than a better fusion.

Table 7: Analytic budget under recursive fusion ($D = 25$, $L = 2$, 30 dB) with the same selector; nothing is retrained. Expert columns spend each condition's budget in one run. Calibrated counts are per expert and were fixed on procedural objects before any test measurement was formed. Gains are computed from unrounded means.

Budget	PhaseFuse-PS	Best expert	Gain	PAP	RAAR	TWF	RAF	ADMM
200 each	35.97	TWF 26.65	+9.32	26.43	20.47	26.65	4.80	16.88
500 each	35.06	TWF 26.47	+8.60	25.72	19.71	26.47	4.80	13.28
Calibrated	**36.03**	RAAR 25.77	+10.26	25.52	25.77	25.63	20.55	24.83

C.3 Synthetic validation versus natural-image test

Table 8 reports the 50-image procedural monitor next to the BSDS500 test for the single-pass reference, routing, and temperature conditions. The two splits are not interchangeable: the procedural gains are larger, and image-level soft routing exceeds per-pixel top-1 routing on BSDS500 but not on the procedural monitor. Neither split selects checkpoints or hyperparameters.

Table 8: Single-pass fusion on the procedural monitor (50 disjoint procedural images, SNR 15–45 dB) and on the BSDS500 test (200 images, 30 dB), 128×128. Best exp. is the score of the best individual expert on each split, and Gain is the fused score minus that value, computed from unrounded means. The monitor uses SNR 15–45 dB and the test 30 dB, so domain and noise effects are confounded.

	Procedural monitor ($N = 50$)			BSDS500 test ($N = 200$)		
Condition	PhaseFuse	Best exp.	Gain	PhaseFuse	Best exp.	Gain
PhaseFuse-PS, $L = 1$	24.338	23.048	$+1.290$	21.174	20.597	$+0.577$
PhaseFuse-PS, $L = 2$	32.039	28.176	$+3.863$	28.696	26.650	$+2.047$
PhaseFuse-PS, $L = 3$	37.529	37.529	≈ 0	37.004	37.004	≈ 0
PhaseFuse-PS, $L = 4$	40.670	40.670	≈ 0	40.172	40.172	≈ 0
Uniform, $L = 2$	23.239	28.176	-4.936	22.859	26.650	-3.791
Image-level blend, $L = 2$	29.710	28.176	$+1.534$	27.929	26.650	$+1.279$
Per-pixel top-1, $L = 2$	30.599	28.176	$+2.423$	27.456	26.650	$+0.807$
$T = 0.3$, $L = 2$	32.194	28.176	$+4.018$	28.766	26.650	$+2.117$
$T = 1.0$, $L = 2$	31.941	28.176	$+3.766$	28.645	26.650	$+1.996$

C.4 Single-pass routing statistics

Table 9 gives the entropy and spatial spread of each routing rule under the single pass on the procedural monitor, and Table 10 gives the mean expert weights of the selectors trained at each temperature. At the reference temperature, TWF receives the largest mean weight, but PAP and the spectral expert retain substantial contributions.

Table 9: Single-pass routing rules ($D = 1$, $\tau = 200$) on the procedural monitor ($L = 2$, SNR 15–45 dB). Phase-PSNR in dB; entropy normalized by $\log 6$; spread in weight units.

Rule	PSNR	Entropy	Spread
Best expert	28.18	—	—
Uniform	23.24	1.000	0.000
Image-level blend	29.71	0.527	0.000
Per-pixel top-1	30.60	0.000	0.218
PhaseFuse-PS	**32.04**	0.485	0.122

Table 10: Mean expert weights (%) and normalized entropy (Ent.) of the selectors trained at each temperature (single pass, procedural monitor, $L = 2$).

T	Spectral	PAP	RAAR	TWF	RAF	ADMM	Ent.
0.3	22.16	30.31	3.96	41.87	0.00	1.71	0.479
0.6	21.55	29.13	4.79	43.62	0.00	0.91	0.485
1.0	22.09	29.08	4.51	43.15	0.30	0.86	0.493

C.5 Training-set size and learning rate

Table 11 reports a separate 64×64 study of all 36 combinations of snapshot count, number of procedural training images, and learning rate, evaluated with the single pass on 50 procedural images at epoch 140. More training images consistently help at 3×10^{-4}, whereas 10^{-3} is often weaker and, for several settings, gives identical scores across training-set sizes because the selector concentrates on a single expert. These scores are diagnostics at a different resolution and are not comparable to the 128×128 results.

Table 11: Single-pass PhaseFuse-PS ($D = 1$, $\tau = 200$) phase-PSNR on 50 procedural images (SNR 15–45 dB) at 64×64 and epoch 140 for each snapshot count, training-set size, and learning rate.

L	Images	10^{-4}	3×10^{-4}	10^{-3}
1	1k	21.730	21.687	21.154
1	5k	22.055	22.090	21.154
1	10k	22.167	22.236	22.249
2	1k	28.834	28.868	26.303
2	5k	29.438	29.476	26.303
2	10k	29.645	29.773	26.303
3	1k	38.530	38.569	35.709
3	5k	39.098	39.152	35.709
3	10k	39.370	39.496	37.402
4	1k	41.394	41.505	40.635
4	5k	41.932	42.024	40.635
4	10k	42.205	42.407	40.635

C.6 Shared versus native initialization

A separate experts-only experiment isolates the effect of initialization on all 200 BSDS500 test images at 64×64, $L = 4$, 30 dB, and 200 updates (Table 12). The native column uses each method's own start: the null initialization for PAP (Algorithm 1 of Chen et al. (2018) with $\gamma = 0.5$, a nonzero vector), the truncated spectral start for TWF, the weighted-correlation start for RAF, and the spectral start for LAD–ADMM; the RAAR release does not prescribe a CDP initializer, so it uses a seeded random-phase start. Under its native start TWF fails, reaching 5.72 dB against 40.15 dB from the shared initialization, which is why all experts share one initialization in the main experiments.

Table 12: Experts-only initialization study (64×64, $L = 4$, 30 dB, 200 BSDS500 test images): final-iterate mean $\pm$ population standard deviation. Parentheses give the iteration with the highest mean over the 200 images and that mean, found with ground truth as a diagnostic only.

Method	Shared initialization	Native initialization
Spectral	17.659 ± 2.431 (0 / 17.659)	17.659 ± 2.431 (0 / 17.659)
PAP	38.045 ± 0.150 (45 / 38.075)	38.043 ± 0.149 (185 / 38.044)
RAAR	27.427 ± 0.576 (10 / 37.009)	27.372 ± 0.577 (55 / 31.433)
TWF	$\mathbf{40.150} \pm 0.139$ (90 / 40.161)	5.715 ± 0.441 (200 / 5.715)
RAF	33.865 ± 6.455 (60 / 36.028)	34.155 ± 6.168 (130 / 34.657)
LAD–ADMM	16.650 ± 0.146 (5 / 30.424)	15.968 ± 0.147 (65 / 16.790)

C.7 Training data and training length

Table 13 and Table 14 give the results discussed in Section 4.5.

Table 13: Training data source under recursive fusion ($D = 25$, $\tau = 8$, $L = 2$, 30 dB; 200 test images per set). The selectors differ only in their training images; FFHQ is unseen by both. Gain is against the strongest expert of each test set, TWF on both (26.65 dB on BSDS500 and 28.67 dB on FFHQ).

Trained on	Tested on	PhaseFuse-PS	Gain
Procedural IFS	BSDS500 test	**35.97**	+9.32
	FFHQ val.	**36.15**	+7.48
BSDS500 train	BSDS500 test	35.48	+8.83
	FFHQ val.	35.45	+6.78

Table 14: Checkpoint epoch for PhaseFuse-PS ($L = 2$, 30 dB, 200 images): single pass ($D = 1$, $\tau = 200$) and recursive ($D = 25$, $\tau = 8$). All epochs were fixed in advance. Hours is the cumulative selector-training time on one RTX 3090, excluding the one-off computation of the expert candidates.

Epoch	Hours	Single	Recursive
20	0.17	28.61	**36.11**
40	0.35	28.67	35.97
80	0.70	28.75	35.90
100	0.87	28.80	36.05
120	1.04	28.76	35.91
140	1.22	28.70	35.97
160	1.39	28.72	35.73
180	1.56	**28.83**	35.71

C.8 Policy gradient over sampled selection

Table 15: REINFORCE fine-tuning of the selector as a per-pixel sampling policy ($L = 2$, 30 dB, 200 images, 20 fine-tuning epochs). Both the warm start and the fine-tuned (RL) selector are read out with the per-pixel blend. W/T/L: images where RL is higher, equal, and lower.

Schedule	Warm	RL	W/T/L
$D = 1$	28.70	23.17	0/0/200
$D = 25$	35.97	22.67	0/0/200
$D = 100$	36.34	35.52	10/0/190

The blend of Eq. (5) is differentiable, so the measurement objective can be minimized directly. The alternative evaluated in Table 15 reads the logits as a per-pixel categorical policy. Let s_d denote the state at decision d, consisting of the aligned candidates, the current estimate, and their residuals, and define

$$\pi_\theta(e \mid s_d, p) = \frac{\exp\{z_e(p)/T\}}{\sum_{k\in\mathcal{E}} \exp\{z_k(p)/T\}}. \tag{18}$$

A sampled reconstruction draws one expert per pixel, $a_p \sim \pi_\theta(\cdot \mid s_d, p)$, and sets $x^{\mathrm{hard}}(p) = \widetilde{v}_{a_p}(p)$, whereas $\boldsymbol{x}^{\mathrm{soft}}$ is the blend of Eq. (5) from the same state. To avoid ground truth, the reward is the measurement objective of Section 3.4 on the second noise realization,

$$J(\boldsymbol{x}) = \frac{\| |\boldsymbol{A}\boldsymbol{x}|^2 - \boldsymbol{y}^{(b)} \|_2^2}{\|\boldsymbol{y}^{(a)}\|_2^2}, \qquad R = -J(\boldsymbol{x}^{\mathrm{hard}}), \tag{19}$$

and the soft blend from the same state serves as baseline, so the advantage $A_d = J(\boldsymbol{x}^{\mathrm{soft}}) - J(\boldsymbol{x}^{\mathrm{hard}})$ is positive when sampling beats blending. The policy is updated by REINFORCE with an entropy bonus,

$$\mathcal{L}_{\mathrm{RL}} = -\mathbb{E}\Big[\, \mathrm{sg}(A_d)\, \tfrac{1}{P} \textstyle\sum_p \log \pi_\theta(a_p \mid s_d, p) \Big] - \beta\, \mathbb{E}\big[\overline{\mathcal{H}}(\pi_\theta)\big], \tag{20}$$

where P is the number of pixels, sg stops the gradient so that the baseline, which depends on θ, contributes none, and $\overline{\mathcal{H}}$ is the per-pixel entropy of π_θ in nats averaged over pixels, with $\beta = 10^{-3}$. The state is that of the single pass ($D = 1$), so each image contributes one decision and one advantage. We warm-start from the trained selector and fine-tune for 20 epochs on the same 10,000 objects with Adam, learning rate 3×10^{-4}, and batch size 128, which gives 1,580 updates. A single scalar

advantage is shared by $P = 128^2$ pixels, so the gradient estimate is noisy, whereas differentiating through the blend gives an exact gradient of the deterministic blending objective, which uses the same measurement criterion but is a different function of θ from the expected loss of sampled selection. The mean advantage moved from -0.25 to -0.31 and the policy entropy rose from 1.72 to 1.79 nats, close to its maximum $\log 6 \approx 1.79$. On the procedural monitor, fine-tuning raised the measurement objective of the blended readout from 0.0088 to 0.0366 and of the sampled readout from 0.0185 to 0.340, while the argmax readout moved from 0.0117 to 0.0142, still above the warm-start blend. These conclusions apply to the tested estimator and hyperparameters. For a schedule with $D > 1$ decisions and a terminal reward $R_{\mathrm{term}} = -J(\boldsymbol{x}_D)$, the corresponding score-function identity is

$$\nabla_\theta \, \mathbb{E}[R_{\mathrm{term}}] = \mathbb{E}\Big[\textstyle\sum_{d=0}^{D-1} \sum_{p=1}^{P} \big(R_{\mathrm{term}} - b_d(s_d)\big) \nabla_\theta \log \pi_\theta(a_{d,p} \mid s_d, p) \Big], \tag{21}$$

assuming an initial state independent of θ, pixel actions that are independent given the state, transitions with no explicit dependence on θ once the actions are fixed, and baselines b_d that do not depend on the actions. With $D = 1$, as in our fine-tuning, it coincides with Eq. (20) up to the $1/P$ normalization and the entropy term.

C.9 Routing behind the qualitative examples

Table 16 gives the routing behind each panel of Figure 3 at the final decision. At $L = 1$, PhaseFuse-GH commits to the spectral expert, i.e., it keeps the current estimate at that decision, and at $L = 3$ and $L = 4$ TWF is the dominant expert at every pixel for both rules.

Table 16: Routing behind each panel of Figure 3 at the final decision ($D = 25$, $\tau = 8$): percentage of pixels whose largest weight is each expert. Entropy is the per-pixel weight entropy in bits averaged over pixels ($\log_2 6 \approx 2.58$ is equal mixing), and Top wt. is the mean over pixels of the largest weight. For the $L = 2$ RGB example the three channels are pooled.

Example	Rule	Entropy	Top wt.	Spectral	PAP	RAAR	TWF	RAF	ADMM
$L = 1$	PhaseFuse-PS	1.60	0.52	46.7	37.1	1.3	14.9	0.0	0.0
	PhaseFuse-GH	0.00	1.00	100.0	0.0	0.0	0.0	0.0	0.0
$L = 2$ (RGB)	PhaseFuse-PS	1.25	0.60	23.2	10.7	0.0	66.2	0.0	0.0
	PhaseFuse-GH	0.00	1.00	0.0	0.0	0.0	100.0	0.0	0.0
$L = 3$	Both	0.00	1.00	0.0	0.0	0.0	100.0	0.0	0.0
$L = 4$	Both	0.00	1.00	0.0	0.0	0.0	100.0	0.0	0.0