

Calibrated per-pixel uncertainty for low-dose paediatric chest-radiograph denoising at no fidelity cost

Samuel Tekle Mengistu

`samuel.mengistu@cs.ox.ac.uk`

University of Oxford

Kyriakos Flouris

University of Cambridge

Research Article

Keywords: image denoising, uncertainty quantification, calibration, variational autoencoder, paediatric chest radiography, low-dose imaging

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10694351/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Calibrated per-pixel uncertainty for low-dose paediatric chest-radiograph denoising at no fidelity cost

Samuel Tekle Mengistu^{*1,2,3,4} and Kyriakos Flouris⁴

¹*Department of Computer Science, University of Oxford, Oxford, United Kingdom*

²*Oxford Martin School, University of Oxford, Oxford, United Kingdom*

³*The Accelerate Hub, University of Oxford, Oxford, United Kingdom*

⁴*MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom*

Abstract

Background. Deep denoisers restore low-dose chest radiographs but emit a single point estimate with no indication of where the output is reliable, so a confident-looking reconstruction can be locally wrong. In paediatric radiography, where the radiation-dose imperative is sharpest, we ask whether a per-pixel uncertainty map can be attached at no meaningful cost to reconstruction quality.

Methods. We develop PP-VAE-Hformer, a hybrid convolution–transformer denoiser with a variational-autoencoder bottleneck (epistemic uncertainty map) and a heteroscedastic dual head (aleatoric map), and run a 19-arm ablation isolating five composite-loss terms on Poisson–Gaussian-degraded paediatric chest radiographs (Kermany collection; 624-image held-out test set), against eight discriminative baselines retrained on identical data and noise. Calibration is assessed by reliability diagrams and post-hoc σ -scaling; comparisons use Bonferroni-corrected Welch tests with Cohen’s d as the primary discriminant.

Results. A heteroscedastic negative-log-likelihood (NLL) objective buys a calibrated aleatoric map for 1.2 dB PSNR, which two-stage fine-tuning recovers to a statistically indistinguishable 0.010 dB. For the variational variant the Kullback–Leibler (KL) annealing *schedule*, not the bottleneck itself, is decisive: cyclic annealing closes a 1.10 dB gap over posterior-collapsing linear warmup. Predicted uncertainty tracks realised error (Pearson $r > 0.8$) with an optimal recalibration scale within 1% of unity; the aleatoric and epistemic maps are only moderately correlated ($\bar{r} = 0.53$), flagging complementary regions. We additionally document a previously unreported destructive interaction between structural-similarity and Sobel-edge supervision at literature-default weights.

*Corresponding author: samuel.mengistu@cs.ox.ac.uk; ORCID: <https://orcid.org/0000-0002-2817-1421>

Conclusions. A calibrated per-pixel uncertainty map can be attached to a paediatric low-dose CXR denoiser at effectively no reconstruction cost. The near-perfect aleatoric calibration partly reflects the matched (“inverse-crime”) noise model, positioning the epistemic map as the genuine reconstruction-uncertainty signal; a blinded reader study and real low-dose data are the immediate next steps.

Keywords. image denoising; uncertainty quantification; calibration; variational autoencoder; paediatric chest radiography; low-dose imaging

1 Background

Chest radiography is the most common paediatric imaging examination, yet the ionising radiation that forms each image carries a measurable lifetime cancer risk that is steeper in children: large cohorts link paediatric imaging radiation to excess leukaemia and brain tumours under a no-safe-threshold model [1–3]. Lowering dose reduces this risk but raises signal-dependent quantum noise, and by the Rose criterion a contrast-to-noise ratio of roughly five is needed before a human observer can reliably detect a low-contrast finding [4, 5]; dose reduction erases faint parenchymal detail long before it touches the high-contrast bony thorax. Every noisy image that forces a repeat acquisition, or an escalation to chest CT at a far greater bone-marrow dose [3], turns that statistical risk into avoidable harm.

Deep denoisers restore much of the lost quality, but from classical filters [6] through discriminative convolutional networks [7, 8] to hybrid transformers [9, 10] they output a single point estimate with no measure of where that estimate can be trusted, while generative denoisers [11, 12] can hallucinate anatomy [13, 14]. None provides a *calibrated* per-pixel uncertainty map that separates pixels recovered from the measurement from pixels filled in by a learned prior. In paediatric radiology—where the dose imperative is sharpest and documented generalisation failures of adult-trained models are a recognised safety concern [15]—this gap is concrete: a denoiser deployed to enable dose reduction should report, spatially, how far its own output can be trusted.

We ask a single question: can replacing the pixel-wise loss with a heteroscedastic, structure- and frequency-aware objective, together with a variational bottleneck, equip a paediatric low-dose chest X-ray (CXR) denoiser with a per-pixel uncertainty map at no clinically meaningful cost to reconstruction quality? We answer it with a controlled 19-arm ablation of a fixed hybrid convolution–transformer backbone, evaluated against eight established discriminative denoisers retrained from scratch on identical data and noise, so that the ablation isolates the training *objective* and the baseline comparison isolates the *architecture*, both free of cross-paper confounds.

Our contributions are:

1. a single 19-arm ablation isolating all five composite-loss terms at once, where prior work tests them in isolation [16, 17], whose ranking holds under held-out perceptual metrics (FSIM, LPIPS) that no arm trains on;
2. the finding that the KL-annealing *schedule*, not the variational bottleneck itself, is the decisive design variable for VAE-based denoising (1.10 dB, cyclic vs. linear warmup), and that two-stage loss-substitution fine-tuning erases the 1.2 dB calibration tax an uncertainty head otherwise costs, to a statistically indistinguishable 0.010 dB;

3. a previously unreported destructive interaction between structural similarity (SSIM) and Sobel-edge supervision at literature-default weights (-1.75 dB, unrecovered by any re-weighting), invisible to either term alone, together with a second guardrail result that PReLU activations destabilise heteroscedastic variance heads; and
4. a calibration analysis showing the aleatoric map is well calibrated at the pixel level (σ -scaling within 1% of unity), read honestly against the inverse-crime caveat [18], with the epistemic map identified as the genuine vehicle for reconstruction uncertainty.

1.1 Related work

Denoising architectures. Learned denoising progressed from residual CNNs [7, 19, 20] to deep noise-conditioned U-Nets [8], windowed-attention transformers [9, 21], gated convolutional designs [22], and Swin-convolution hybrids trained on realistic noise [23]; Hformer [10] interleaves convolution and attention for low-dose CT. For paediatric CXR specifically, SharpXR [24] pairs a denoising decoder with a Laplacian edge branch. All are deterministic: none attaches a reliability signal to its output.

Loss functions. Pixel norms (L2, L1, Charbonnier [25]) converge to the posterior mean and over-smooth [16]; structural [26, 27] and frequency-domain [28, 29] terms recover perceptual and spectral content that spectral bias suppresses [30]. Prior work evaluates these terms largely in isolation; interactions between them at literature-default weights are undocumented.

Uncertainty quantification. Heteroscedastic regression heads capture aleatoric (data) uncertainty and Monte-Carlo sampling of a stochastic representation captures epistemic (model) uncertainty [31, 32]; deep ensembles offer an architecture-independent alternative [33]. Variational autoencoders [34] supply a principled stochastic bottleneck but are prone to posterior collapse, mitigated by cyclical annealing [35, 36] or free bits [37]. A 2025 survey of VAE applications in medical imaging [38] records no system combining hybrid-attention reconstruction with calibrated per-pixel aleatoric and Monte-Carlo epistemic maps on paediatric CXR.

2 Methods

2.1 Study design

Nineteen arms were defined on a single fixed backbone: sixteen core arms (A–P) each differing only in the loss function or one specific design choice (probabilistic vs. deterministic latent space, KL schedule, activation function), plus a Charbonnier pixel-norm variant (Q) trained from scratch and two fine-tuning arms (R, S) that resume from a converged checkpoint. Each arm adds exactly one component to its predecessor, so performance differences are attributable to the component under test [39]. Every model—the 19 arms and eight retrained baselines—is trained and evaluated on identical data and noise. Table 1 summarises the design; Fig. 1 shows the end-to-end pipeline.

Table 1: Active loss terms and design features for all 19 ablation arms. All other settings (architecture, optimiser, learning rate, data, schedule) are identical across arms.

Arm	Objective / design variable	VAE
A	MSE baseline (replicates DnCNN/Hformer training)	–
B	Heteroscedastic NLL	–
C	NLL + MS-SSIM	–
D	NLL + MS-SSIM + FFL	–
E	NLL + SSIM + FFL + KL, linear warmup	✓
F	as E, cyclical KL annealing	✓
G	as E, free-bits KL	✓
H	as E, cyclical + free-bits KL	✓
I	L1	–
J	L1 + SSIM + FFL	–
K	NLL + L1	–
L	NLL + Sobel edge + FFL	–
M	NLL + SSIM + Sobel edge + FFL	–
N	NLL + VGG-16 perceptual + SSIM + FFL	–
O	as D, PReLU decoder activations	–
P	as H + edge, PReLU activations	✓
Q	Charbonnier pixel norm, from scratch	–
R	two-stage FT from A: L1+SSIM+FFL	–
S	two-stage FT from A: NLL+SSIM+FFL	–

2.2 Poisson–Gaussian degradation

We simulate low dose with the Foi affine-variance model [40], $\text{Var}(\tilde{y} \mid y) = ay + b$, applied on-the-fly to displayed intensity $y \in [0, 1]$ using a Gaussian approximation to the Poisson term, $\tilde{y} = y + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, a \max(y, 0) + b)$, clipped to $[0, 1]$. Three presets probe mild, moderate and severe noise ($(a, b) = (0.01, 0.002), (0.03, 0.005), (0.08, 0.010)$; effective $\sigma \approx 15, 25, 50$ at mean intensity, matching the standard evaluation grid [7]). Training noise is freshly sampled each epoch with the preset drawn uniformly per batch; validation and test noise is fixed per deterministic seed. The implementation was validated against its specification: 4×10^4 realisations at each of 51 intensity levels recover the specified (a, b) to within 1% with $R^2 > 0.999$ at every preset.

Applying the affine law in the image domain is standard for processed radiographs [24, 41] but, because displayed brightness is inverted relative to photon flux, places noise on the bright bony thorax rather than the photon-starved lung. The noise remains signal-dependent and identical for every arm, so within-study comparisons are unaffected; the choice is this study’s principal degradation limitation (Section 4.1).

2.3 Architecture

PP-VAE-Hformer keeps the reconstruction backbone of Hformer [10]—a hybrid U-Net whose stages interleave a residual convolution branch with self-attention—realising the attention branch as non-overlapping 8×8 windowed self-attention following Swin [21], and adds the two components that make the output probabilistic. A 3×3 stem maps the

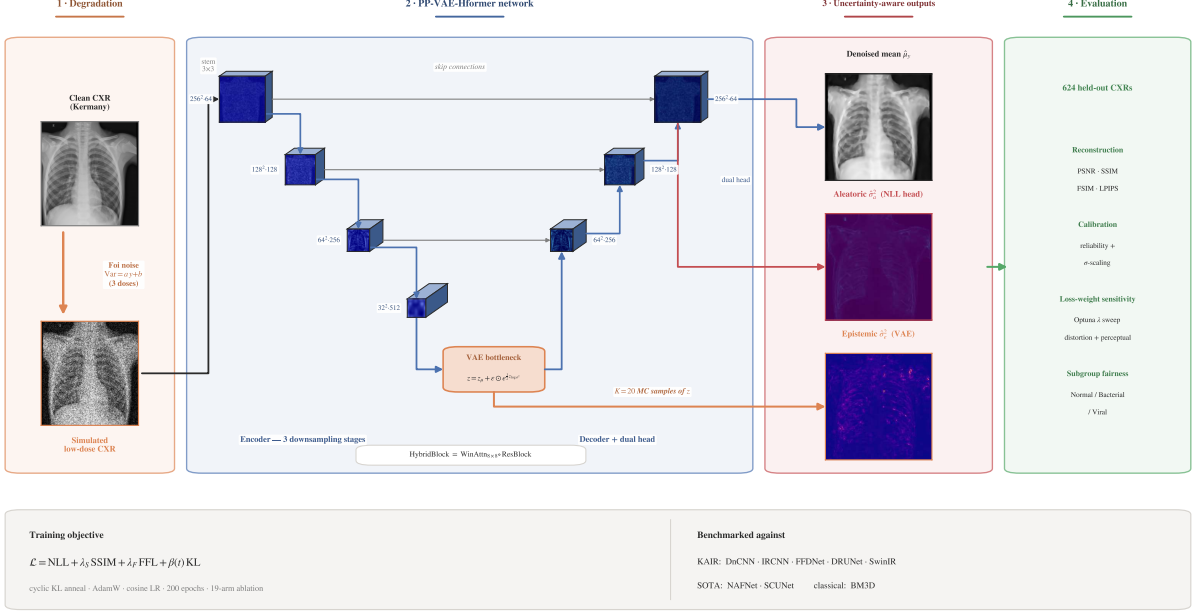


Figure 1: End-to-end study pipeline. (1) A clean Kermany CXR is degraded by the Foi noise model at three dose levels. (2) The noisy image passes through the PP-VAE-Hformer U-Net with an optional VAE bottleneck. (3) The dual head outputs the denoised mean $\hat{\mu}_y$ and aleatoric variance $\hat{\sigma}_a^2$; the epistemic variance $\hat{\sigma}_e^2$ comes from $K=20$ Monte-Carlo samples. (4) Evaluation is on 624 held-out CXRs against eight retrained baselines.

$1 \times 256 \times 256$ input to 64 channels; three encoder stages of two HybridBlocks each (ResBlock with GroupNorm [42] and GELU [43], residual skip [44], followed by windowed multi-head attention [45], $h = 4$) halve resolution and double channels to a $512 \times 32 \times 32$ bottleneck; the decoder mirrors this with transposed convolutions and fused skip connections. Windowed attention keeps cost linear in pixels and the network compact at 25.7 M parameters (24.9 M without the VAE bottleneck), well below several baselines. All weights are trained from scratch; no pretraining or transfer learning is used.

VAE bottleneck and epistemic map. For the variational arms, three 1×1 convolutions at the bottleneck produce the posterior mean $\mathbf{z}_\mu$, log-variance $\mathbf{z}_{\log \sigma^2}$, and the projection of the reparameterised sample $\mathbf{z} = \mathbf{z}_\mu + \boldsymbol{\varepsilon} \odot \exp(\frac{1}{2} \mathbf{z}_{\log \sigma^2})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ [34]. At evaluation, reconstruction uses the deterministic pass $\mathbf{z} = \mathbf{z}_\mu$; the epistemic map is the pixel-wise variance of $K=20$ stochastic passes, which under the law of total variance complements the aleatoric term [32].

Dual output head. Two parallel 3×3 heads at full resolution emit the reconstruction mean $\hat{\mu}_y$ and the per-pixel log aleatoric variance $\log \hat{\sigma}_a^2$ in a single forward pass.

2.4 Composite objective

The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \beta_t \mathcal{L}_{\text{KL}} + \lambda_{\text{SSIM}} \mathcal{L}_{\text{SSIM}} + \lambda_{\text{FFL}} \mathcal{L}_{\text{FFL}} [+ \lambda_{\text{edge}} \mathcal{L}_{\text{edge}}], \quad (1)$$

with heteroscedastic NLL [31]

$$\mathcal{L}_{\text{NLL}} = \frac{1}{HW} \sum_i \left[\frac{(y(i) - \hat{\mu}_y(i))^2}{2 \exp \hat{s}(i)} + \frac{\hat{s}(i)}{2} \right], \quad \hat{s}(i) = \log \hat{\sigma}_a^2(i), \quad (2)$$

multi-scale SSIM [27] ($\lambda_{\text{SSIM}} = 0.5$ [16]), focal frequency loss [28] ($\lambda_{\text{FFL}} = 0.1$, $\gamma = 1$), an optional Sobel-gradient L1 edge term in the sense of Mathieu et al. [46] ($\lambda_{\text{edge}} = 0.1$; Arms L, M, N, P), and the closed-form Gaussian KL divergence [47]. Arm E anneals β linearly to $\beta_{\text{max}} = 10^{-3}$ over 20 epochs; Arm F uses cyclical cosine annealing over four 50-epoch cycles [35]; Arm G clips per-channel KL below $\lambda = 0.25$ nats (free bits [37]); Arm H combines both.

Two-stage loss-substitution fine-tuning. Arms R and S load Arm A’s converged L2 checkpoint and substitute a richer composite loss for 100 further epochs (R: L1+SSIM+FFL, $\eta = 5 \times 10^{-5}$; S: NLL+SSIM+FFL, $\eta = 2 \times 10^{-5}$, with a 20-epoch linear loss blend to stabilise the variance head), testing whether the NLL fidelity cost is recoverable without retraining.

2.5 Data

The Kermany paediatric CXR collection [48] comprises 5,856 anterior–posterior radiographs from children aged one to five years. Images were converted to greyscale, resized to 256×256 with antialiased bilinear interpolation, and normalised to $[0, 1]$. The 624-image test partition (234 Normal, 242 Bacterial, 148 Viral Pneumonia) was locked before development; of the remainder, 785 images (15%) formed the validation set (checkpoint selection only) and 4,447 the training set. Spatial augmentation (horizontal flip, rotation $\pm 5^\circ$) was applied to training only; intensity augmentation was excluded as it would invalidate the calibrated noise parameters. Class labels were unused during training.

2.6 Baselines

Eight established discriminative denoisers were retrained from scratch on the identical split and noise model: DnCNN [7], IRCNN [19], FFDNet [20], DRUNet [8], SwinIR [9], NAFNet [22], SCUNet [23], and the paediatric-CXR-specific SharpXR [24] (under its authors’ RMSE objective); classical BM3D [6] with oracle noise knowledge completes the set. The first seven share the ablation’s training protocol. Because every baseline is retrained within-study, absolute figures are directly comparable to each other rather than to the original papers.

2.7 Training

All arms trained for 200 epochs, batch size 16, AdamW [49] at 2×10^{-4} with cosine annealing to 10^{-6} , weight decay 10^{-4} , mixed precision [50], and gradient clipping at 1.0, on NVIDIA A100 80 GB GPUs (Cambridge CSD3; ≈ 9 – 10 h per proposed arm, ≈ 398 GPU-hours total). The best validation-PSNR checkpoint was selected per arm.

2.8 Evaluation and statistical analysis

Reconstruction is scored by PSNR, SSIM and MS-SSIM; the held-out FSIM [51] (primary) and LPIPS [52] appear in no training objective and serve as unbiased perceptual judges. Calibration uses reliability diagrams (test pixels binned by predicted $\hat{\sigma}_a$ against realised RMSE, summarised by Pearson r), NLL, sharpness, and a post-hoc σ -scaling factor s^* fitted on held-out pixels [53]. Values are means over the 624-image test set with 95% bootstrap confidence intervals (1,000 resamples) [54].

Because the central claims are claims of *equivalence*, the analysis is built around effect magnitude. An omnibus ANOVA gates $\binom{24}{2} = 276$ pairwise Welch t -tests [55] (robust to the strong heteroscedasticity between stable and failed arms), Bonferroni-corrected to $\alpha_{\text{adj}} = 1.81 \times 10^{-4}$ [56]. At $n = 624$ nearly every pair reaches nominal significance [57], so Cohen’s d [58] is the primary discriminant ($|d| < 0.2$ negligible). Subgroup fairness across diagnostic classes uses Welch ANOVA with Bonferroni-adjusted Games–Howell tests. Analyses used PyTorch [59], SciPy [60], Optuna [61] for the loss-weight sweep, and R for inferential statistics.

3 Results

Three results carry this section. First, the uncertainty map has a price, and it is small: the heteroscedastic likelihood buys the per-pixel aleatoric map for about 1.2 dB of PSNR. Second, for the VAE variant the KL-annealing schedule is decisive: cyclic annealing recovers a 1.10 dB gap over posterior-collapsing linear warmup, reaching statistical parity with the retrained CNN baselines. Third, the 1.2 dB price is not fundamental: two-stage fine-tuning closes it to a statistically indistinguishable 0.010 dB.

3.1 Reconstruction performance

Table 2 reports mid-noise reconstruction for the key arms and baselines. Because each test image carries fixed noise and each model returns a single reconstruction, the metrics are exactly repeatable, so sub-decibel gaps are genuine differences.

The pixel-norm and fine-tuned arms hold the reconstruction ceiling (Arm A 34.62 dB, Arm S 34.61 dB, Arm R 34.55 dB), edging every retrained baseline: the strongest, SCUNet, reaches 34.15 dB despite its 57 M parameters, with NAFNet (116 M) and SharpXR further back. Adding the heteroscedastic NLL head costs about 1.2 dB (Arm B vs. A)—the central price the rest of this section tests. The VAE bottleneck under cyclic KL (Arm H) trades a further 0.7 dB for epistemic uncertainty; the collapsed linear-KL Arm E and the PReLU failure Arm O mark the floor, and every trained arm except Arm O beats BM3D even with oracle noise knowledge, confirming that additive-white-Gaussian estimators are structurally inadequate for Poisson–Gaussian noise.

3.2 Held-out perceptual metrics

Neither FSIM nor LPIPS appears in any training objective, so both are fully held-out judges (Table 2, right columns). The broad ranking is preserved (rank correlation with PSNR $r_s = 0.95$), with one decisive exception: the perceptual optimum is not the PSNR optimum. Arm R posts the lowest LPIPS (0.2192) of any condition and the highest FSIM of any ablation arm (0.9332), while ranking only third on PSNR (-0.073 dB vs. Arm A).

Table 2: Reconstruction and held-out perceptual metrics at mid noise ($n=624$, mean; PSNR $\pm$ SD). Best value per column in bold. Unc.: uncertainty maps produced (Ale. aleatoric, Epi. epistemic).

Model	Unc.	PSNR (dB)	SSIM	FSIM $\uparrow$	LPIPS $\downarrow$
<i>Pixel-norm and fine-tuned arms</i>					
Arm A (L2)	—	34.618 ± 1.071	0.9269	0.9317	0.2257
Arm R (FT: L1+SSIM+FFL)	—	34.545 ± 1.067	0.9269	0.9332	0.2192
Arm S (FT: NLL+SSIM+FFL)	Ale.	34.608 ± 1.076	0.9266	0.9317	0.2260
Arm Q (Charbonnier)	—	34.535	—	0.9315	0.2242
Arm I (L1)	—	34.444	—	0.9296	0.2291
<i>Uncertainty arms (single-stage)</i>					
Arm B (NLL)	Ale.	33.46	0.909	0.9157	0.2732
Arm D (NLL+SSIM+FFL)	Ale.	33.435 ± 1.026	0.9089	0.9160	0.2733
Arm H (PP-VAE, cyc+fb)	Ale.+Epi.	32.762 ± 1.064	0.8957	0.9056	0.3079
Arm F (PP-VAE, cyclic)	Ale.+Epi.	32.73	—	0.9072	0.3009
Arm E (PP-VAE, linear; collapsed)	Ale.	31.635 ± 0.951	0.8726	0.8926	0.3687
<i>Failure mode</i>					
Arm O (PReLU)	—	29.074 ± 0.909	0.8037	0.8757	0.4397
<i>Retrained baselines (no uncertainty map)</i>					
SCUNet	—	34.149 ± 1.094	0.9127	0.9335	0.2201
NAFNet	—	33.803 ± 1.100	0.9068	0.9288	0.2305
SharpXR	—	33.561 ± 1.107	0.9019	0.9249	0.2368
DRUNet	—	32.58	0.866	0.9212	0.2354
DnCNN	—	32.798 ± 0.941	0.8774	0.9071	0.2922
SwinIR	—	31.18	0.853	0.8905	0.4148
BM3D (oracle σ)	—	30.716 ± 0.929	0.8056	0.8916	0.3716

One condition sits above it on FSIM—SCUNet (0.9335), a significant but negligible-sized gap ($d = 0.19$)—which SCUNet concedes on LPIPS ($d = 0.22$); the two split the perceptual metrics. Arm R’s single-stage parent Arm J is inferior on both, so the advantage comes specifically from the two-stage schedule: an L1-trained checkpoint fine-tuned under a composite perceptual objective. This is the perception–distortion trade-off [62] resolved in the perceptual direction.

All conditions degrade monotonically across the three noise presets (Fig. 2), and the ranking is stable across the dose range (Spearman $r_s > 0.98$), so mid-noise comparisons are representative rather than an artefact of one operating point. Fig. 3 shows all 27 conditions across all four metrics at once.

3.3 Statistical comparisons

The Bonferroni-corrected pairwise tests confirm six conclusions. (i) L1, L2 and Charbonnier are statistically indistinguishable in PSNR, SSIM and LPIPS ($d < 0.2$), confirming the Bayes-optimal-estimator equivalence of pixel norms under approximately Gaussian marginals [16]. (ii) NLL alone and NLL+SSIM+FFL are indistinguishable in PSNR, but adding SSIM+Edge (Arm M) causes catastrophic gradient interference. (iii) Cyclic KL outperforms linear warmup by 1.10 dB; free bits added to cyclic is equivalent to cyclic alone. (iv) The VAE bottleneck itself imposes no detectable PSNR penalty: Arms E and F are

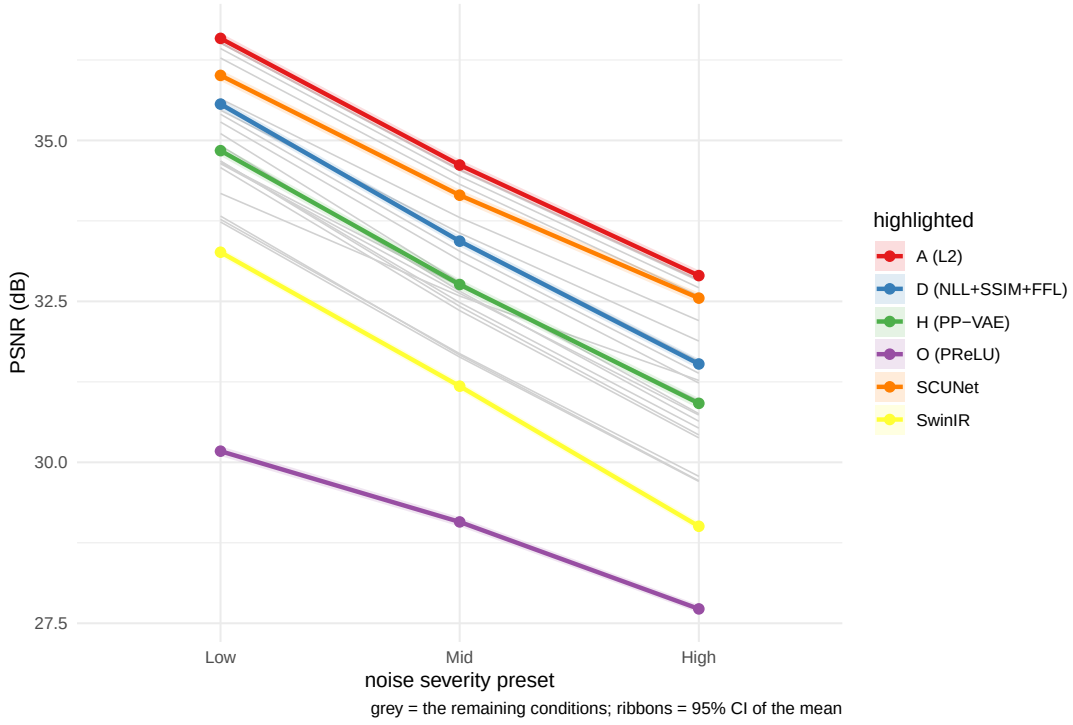


Figure 2: Mean PSNR versus noise-severity preset for all 27 conditions (19 arms and eight retrained baselines). Arm A leads at every dose ahead of SCUNet, the strongest retrained baseline; Arms D and H follow $\approx 1\text{--}2$ dB below with their uncertainty maps; SwinIR trails the CNN baselines; Arm O (PReLU) collapses at every level. Six highlighted conditions carry 95% CI ribbons ($n=624$).

architecturally identical, so the 1.10 dB gap reflects the schedule alone. (v) PReLU collapses catastrophically ($d = +5.58$ vs. Arm D). (vi) In PSNR the KAIR CNN baselines are indistinguishable from the cyclic-KL VAE arms, so the VAE adds calibrated uncertainty at no measurable PSNR cost relative to that family; the retrained SOTA baselines (NAFNet, SCUNet, SharpXR) form a higher cluster, still significantly below the pixel-norm and fine-tuned arms.

3.4 Recovering the cost: two-stage fine-tuning

Table 3 and Fig. 4 summarise the two-stage loss-substitution arms. Neither fine-tuned arm is separable from the Arm A base (S vs. A: $t = 0.17$, $p = 0.87$, $d = 0.01$), while two-stage beats single-stage-from-scratch decisively (S vs. D: $t = 19.7$, $p < 10^{-70}$, $d = 1.11$). Arm S therefore delivers the calibrated aleatoric map (NLL -2.650 , sharpness 0.0201) at the L2 ceiling: the 1.2 dB calibration tax is a training-trajectory artefact, not a fixed cost.

3.5 Qualitative and region-of-interest assessment

Fig. 5 shows Arm H on one representative case per diagnostic class. The denoised mean restores the anatomy the noise hid; the aleatoric map is bright along rib, vessel and consolidation borders and dark over smooth lung; and the $\hat{\sigma}_a$ and error columns light up in the same places—the visual form of the pixel-level calibration quantified below. The epistemic column flags a partly different set of regions (consolidation and cardiac borders),

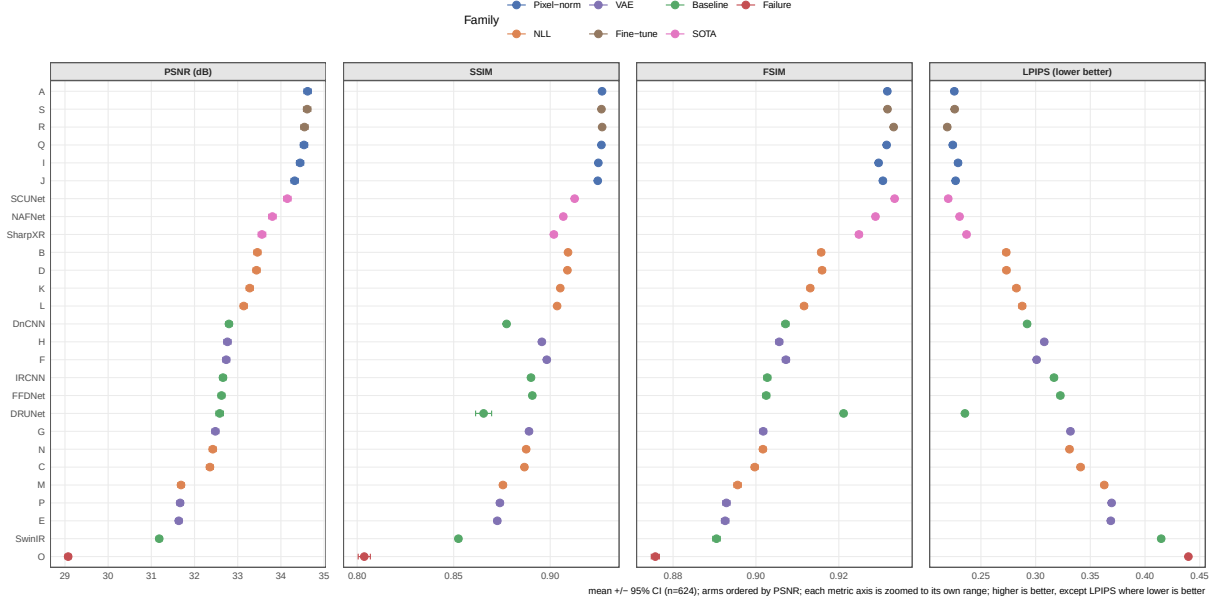


Figure 3: All 27 conditions across every metric at mid noise, ordered by PSNR and coloured by loss/architecture family (PSNR, SSIM, FSIM higher is better; LPIPS lower). The held-out FSIM and LPIPS preserve the family-level block structure of the PSNR ranking while reordering arms within blocks, so the broad ranking is not an artefact of any single objective. Error bars are 95% CIs ($n=624$).

Table 3: Two-stage loss-substitution fine-tuning vs. single-stage counterparts at mid noise.

Arm	Objective	PSNR (dB)	LPIPS↓	Δ PSNR vs. A
A	L2 (Stage-1 base)	34.618 ± 1.071	0.2257	—
R	L1+SSIM+FFL (Stage 2)	34.545 ± 1.067	0.2192	−0.073 dB
S	NLL+SSIM+FFL (Stage 2)	34.608 ± 1.076	0.2260	−0.010 dB
J	L1+SSIM+FFL (scratch)	34.319 ± 1.060	0.2268	−0.299 dB
D	NLL+SSIM+FFL (scratch)	33.435 ± 1.026	0.2733	−1.183 dB

and exists only because the cyclic schedule keeps the latent active.

Region-of-interest crops at the four hardest, most diagnostically decisive sites (fine pulmonary vessels, posterior rib cortical margin, costophrenic angle, perihilar vessels; Fig. 6) show that *no* condition—ours or the far larger SCUNet and NAFNet—resolves any site back to the reference; the objective sets only how soft the residual looks. The frequency-aware arms retain high-frequency hilar and vascular texture that mean-seeking losses smooth away, which is why Arm R attains the best LPIPS at a sub-decibel PSNR cost. SwinIR produces a faint periodic grid at all four sites, most concerning at the costophrenic angle where it could mimic early effusion blunting.

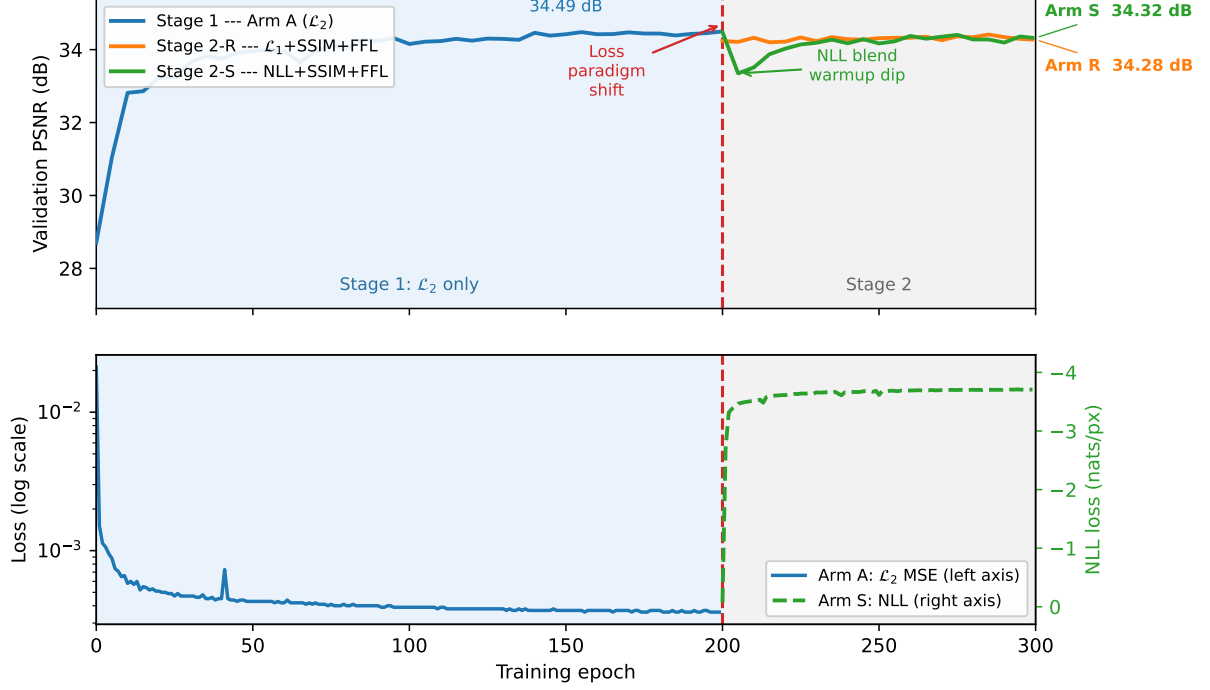


Figure 4: Two-stage fine-tuning dynamics. Top: 300-epoch validation-PSNR trajectory—Arm A converges on MSE (epochs 0–199), then Arms R and S resume from its weights under a composite loss (dashed line at epoch 200); Arm R holds Arm A’s level throughout, while Arm S dips as its variance head re-stabilises and recovers within ≈ 40 epochs. Bottom: corresponding training losses (MSE left axis, NLL right axis).

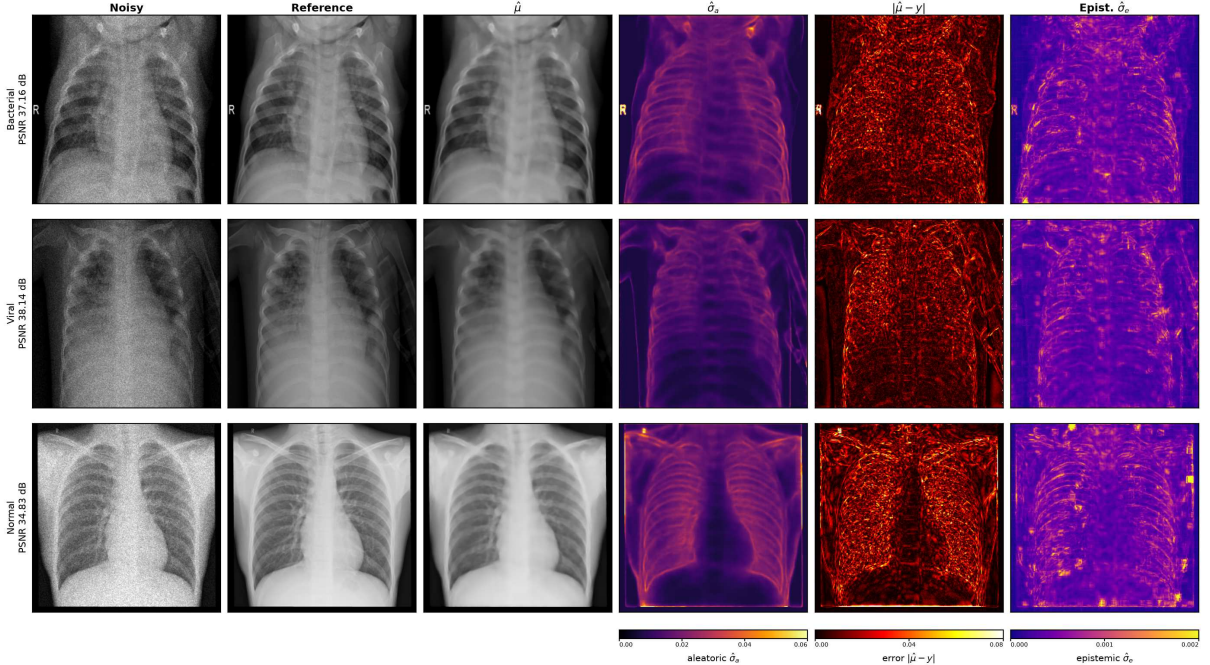


Figure 5: Arm H (best VAE) on one representative case per class at mid noise. Columns: noisy input, clean reference, denoised mean $\hat{\mu}$, aleatoric map $\hat{\sigma}_a$ (inferno, shared 0–0.06 scale), absolute error $|\hat{\mu} - y|$, epistemic map $\hat{\sigma}_e$ (plasma, own scale). Epistemic uncertainty localises to consolidation boundaries (Bacterial) and cardiomegaly margins (Viral), staying low in Normal cases.

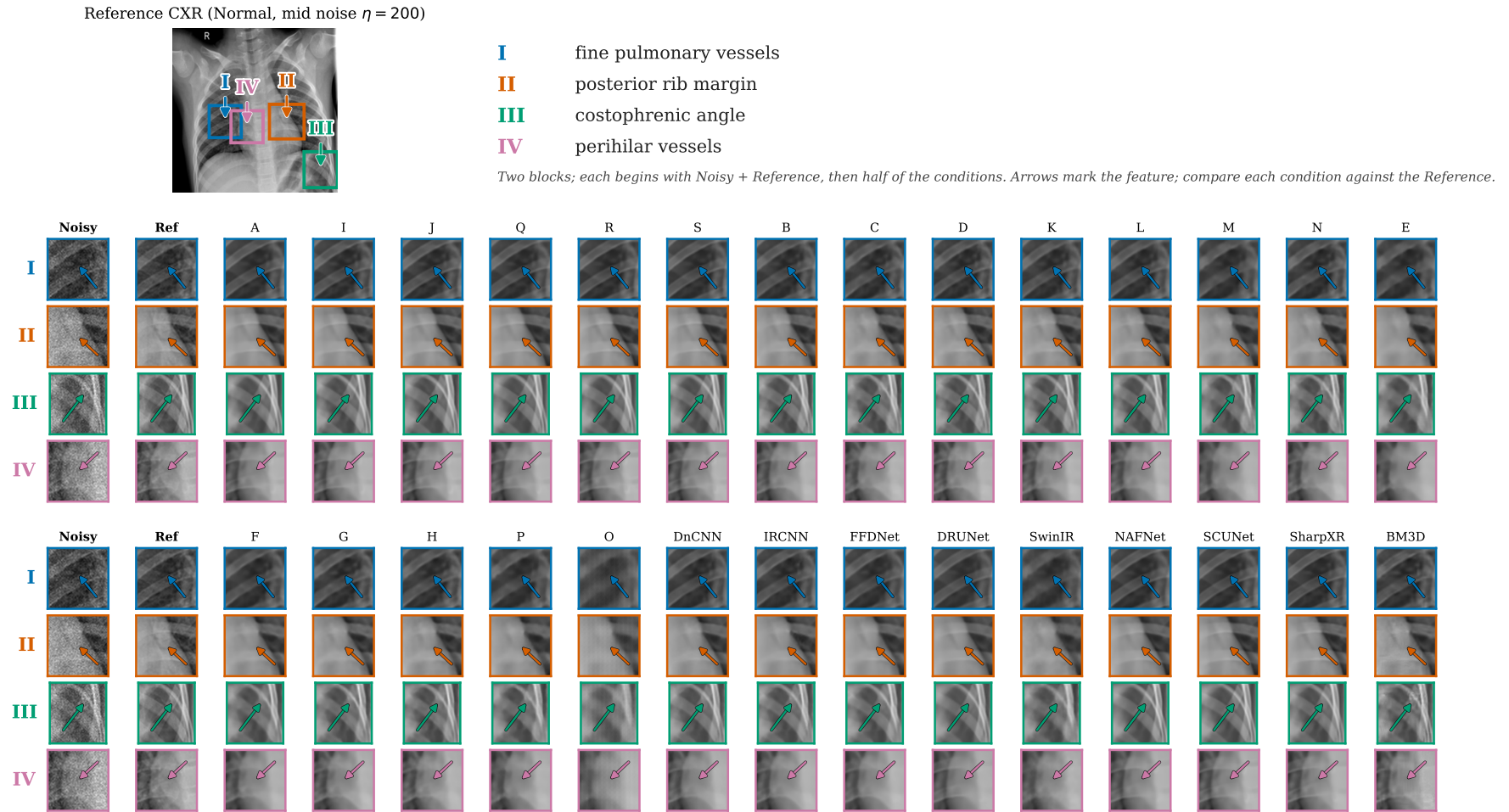


Figure 6: Region-of-interest comparison at mid noise. Top left: reference CXR with four arrowed sites (I fine vessels, II posterior rib margin, III costophrenic angle, IV perihilar vessels). Below: zoomed crops for every ablation arm and all eight retrained baselines. No model recovers the finest vessels to full realism; differences are of degree, not kind.

3.6 Sensitivity to loss weights

A 14-trial Optuna sweep of Arm D’s composite weights finds a smooth, single-peaked validation-PSNR optimum ($\lambda_{\text{SSIM}} \approx 0.96$, $\lambda_{\text{FFL}} \approx 0.02$) about 1.2 dB above the literature-default weights used throughout, so the reported ablation metrics are conservative; all four metrics agree on the optimum (Spearman $r_s \geq 0.89$). A supplementary edge/SSIM sweep finds no re-weighting that recovers Arm M.

3.7 Latent space

The bottleneck is the source of the epistemic map, so whether its KL schedule trained the latent rather than collapsing it underwrites that signal. k -means clustering of the pooled posterior means ($n = 448$ test images, $k = 3$) shows the schedule is the primary driver of class structure: Arm F (cyclical) reaches adjusted Rand index 0.333 [0.242, 0.429] and silhouette 0.113 with visible class regions in t-SNE/UMAP projections (Fig. 7), Arm H (cyc+fb) 0.210, while the collapsed Arm E gives ARI -0.002 and free-bits-only 0.048. Arm F separates classes better while Arm H reconstructs marginally better, a structure-versus-reconstruction choice rather than a single best setting.

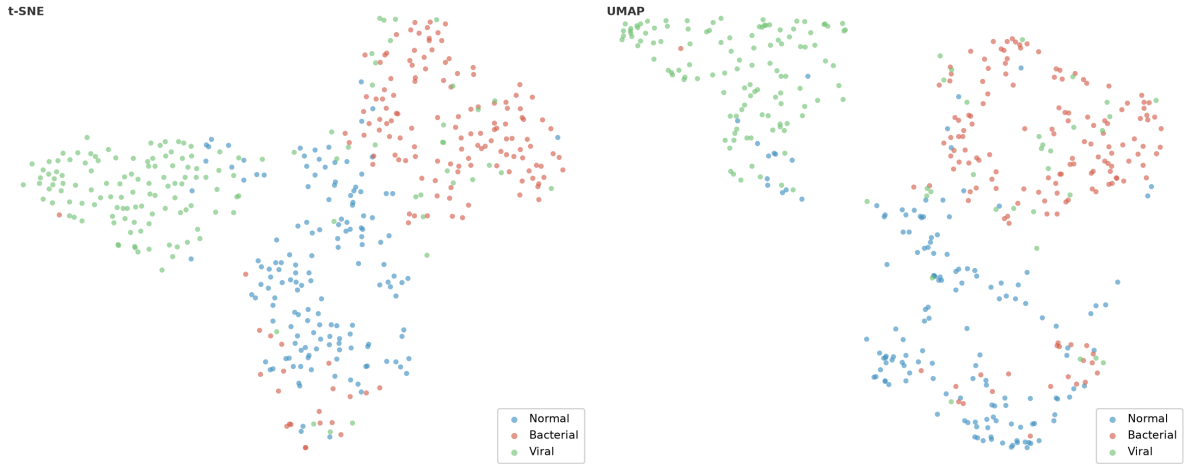


Figure 7: t-SNE (left) and UMAP (right) of Arm F latent codes ($n=448$), coloured by ground-truth class. Class-structured regions emerge although the model sees no class labels (ARI = 0.333, the highest of the five VAE arms); Viral forms the most compact cluster, Normal and Bacterial partially separate.

3.8 Calibration

Arms B and D achieve NLL ≈ -2.64 nats/pixel at mid noise (Arm H ≈ -2.55), sharpness 0.020 (B/D) and 0.022 (H). Reliability diagrams (Fig. 8) show observed RMS error tracking predicted $\hat{\sigma}_a$ along the diagonal (Pearson $r \approx 0.8$ – 0.9), and the post-hoc σ -scaling puts the size on a formal footing: the optimal scale is within 1% of unity ($s^* = 1.005, 1.008, 1.002$ for Arms B, D, H) with mean absolute calibration error ≈ 0.001 , essentially unchanged by rescaling. A predicted standard deviation can therefore be read close to face value: a pixel flagged $\hat{\sigma} \approx 0.02$ carries an error of about 0.02, not merely more than one flagged 0.01. The map also behaves as a dose-response signal: mean $\tilde{\sigma}_a$ rises monotonically as dose falls (Arm D $0.020 \rightarrow 0.031$; Arm H $0.022 \rightarrow 0.034$), with Arm H’s epistemic map

climbing in parallel ($0.0006 \rightarrow 0.0011$), consistent with the model leaning more on its prior as the measurement weakens.

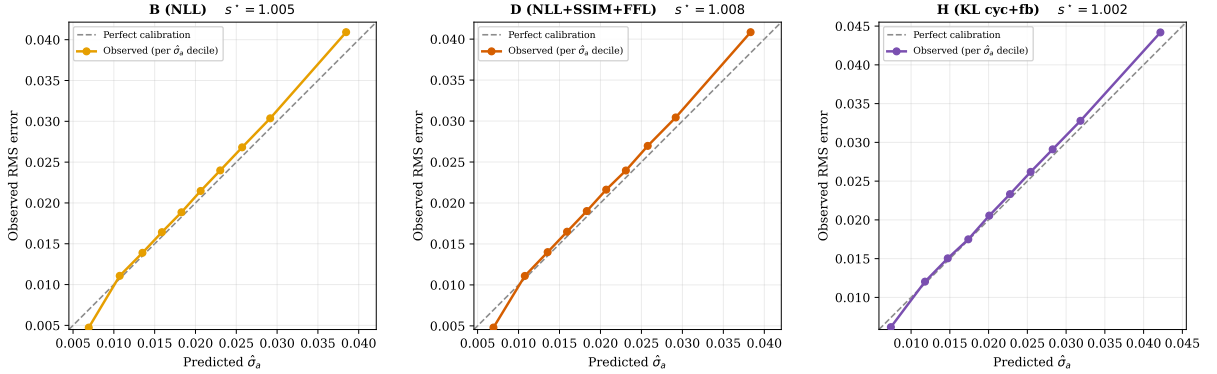


Figure 8: Pixel-level reliability diagrams for Arms B, D and H at mid noise: held-out test pixels grouped into deciles of predicted $\hat{\sigma}_a$, each decile’s observed RMS error plotted against its mean prediction; the dashed diagonal is perfect calibration. All three track the diagonal closely, with the single post-hoc scale s^* within 1% of unity (annotated per panel). The inverse-crime caveat of the Discussion applies.

Aleatoric and epistemic maps are complementary. For Arm H the aleatoric map is $\sim 30\times$ larger in magnitude and concentrates on the high-contrast bony thorax, consistent with re-expressing the injected Foi noise; the epistemic map is only moderately correlated with it (per-image Pearson $\bar{r} = 0.53 \pm 0.07$ over 100 images; top-decile Jaccard 0.21 ± 0.03 ; Fig. 9). The epistemic map is not a rescaled copy: the aleatoric channel answers “how noisy is this pixel” and the epistemic channel “how much of this pixel is invented by the model.”

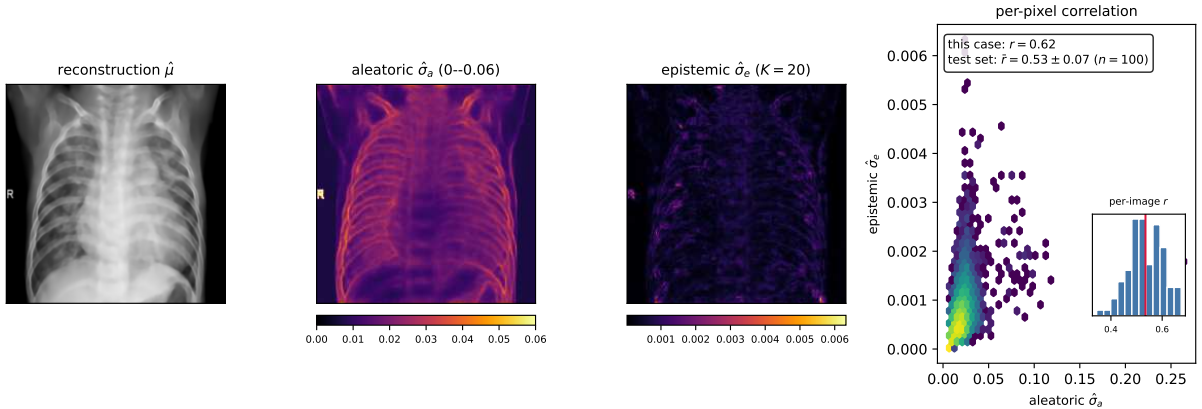


Figure 9: Aleatoric and epistemic uncertainty are complementary (Arm H, one Bacterial case). Left to right: reconstruction $\hat{\mu}$; aleatoric $\hat{\sigma}_a$ (shared 0–0.06 scale, concentrated on the bony thorax and edges); epistemic $\hat{\sigma}_e$ ($K=20$ samples, own scale, $\sim 30\times$ smaller and spatially distinct); per-pixel correlation panel (this case $r = 0.62$; inset: distribution of per-image r over 100 images, $\bar{r} = 0.53 \pm 0.07$).

3.9 Fairness across diagnoses

No objective is biased against the pathological classes: both pneumonia classes reconstruct at higher PSNR than Normal in every model, by a nearly constant margin (Bacterial +1.1, Viral +0.9 dB), because lobar consolidation replaces fine vascular detail with smooth opacity that is easier to denoise; the ordering is a property of the images, not the models, and the held-out LPIPS agrees.

3.10 Failure modes as design lessons

Three arms fail instructively (Table 4), and share a lesson about heteroscedastic training: PReLU’s learnable negative slope lets the log-variance head diverge (variance-predicting layers need bounded activations such as GELU); linear KL warmup collapses the posterior to a deterministic decoder with structureless epistemic maps; and SSIM and Sobel-edge terms pressure the same spatial gradients through conflicting pathways—a previously unreported interaction invisible to either term alone and unrecovered by any re-weighting.

Table 4: Failure modes with mechanism and PSNR impact at mid noise (vs. Arm D).

Arm	Failure	Mechanism	PSNR
O	PReLU instability	log-variance head diverges (epoch 12)	29.07 dB (−4.36)
E	Posterior collapse	linear KL warmup; $D_{\text{KL}} < 0.01$ nats	31.64 dB (−1.80)
M	Gradient interference	SSIM + Sobel supervise the same gradients	31.69 dB (−1.75)

4 Discussion

The answer to the research question is yes. The calibrated per-pixel uncertainty map is obtained at effectively no cost to reconstruction quality. The heteroscedastic head’s 1.2 dB price is a training-trajectory artefact, not a ceiling: two-stage fine-tuning closes it to 0.010 dB, statistically indistinguishable from the L2 base. The same fine-tuning resolves the perception–distortion trade-off [62] in the perceptual direction, with Arm R and SCUNet splitting the two held-out perceptual judges without either leading PSNR.

Two levers matter more than parameter count. Among the retrained CNN baselines, years of architectural change buy little (DnCNN through DRUNet sit within 0.22 dB of one another). The first lever is the backbone: the hybrid convolution–attention design exceeds DnCNN by 1.82 dB at the identical L2 loss, at 25.7 M parameters against SCUNet’s 57 M and NAFNet’s 116 M, so efficient attention rather than scale sets the ceiling. The second lever is the objective on that fixed backbone: the KL schedule alone is worth 1.10 dB, the SSIM+FFL terms buy perceptual quality a pixel loss discards, and the NLL head buys the calibrated map at a cost fine-tuning erases. The NLL+SSIM+FFL objective is architecture-agnostic: it requires only a (mean, log-variance) decoder head and transfers to any heteroscedastic encoder–decoder. A cautionary corollary: pure windowed self-attention (SwinIR), competitive on additive-Gaussian benchmarks, trails plain CNNs under Poisson–Gaussian paediatric CXR noise.

An information limit, not a capacity limit, bounds every denoiser. At the diagnostically decisive regions no model recovers the finest vessels to full realism: once low-dose noise buries a thread-like vessel, the data-processing inequality guarantees no deterministic map can restore what the measurement did not preserve [63]. What a denoiser does next is the forced choice of the perception–distortion trade-off: a mean-seeking objective smooths the missing detail into blur [16], whereas a generative denoiser [11, 12, 64] paints in a sharp, plausible hallucination [13, 14]—arguably the more dangerous failure for a diagnostic image. A calibrated uncertainty map is precisely the instrument that makes this limit visible per pixel.

What the uncertainty map is, and is not. The map’s meaning must be read with care. Because the degradation is simulated with the Foi model and the heteroscedastic head is trained to predict $ay + b$, the aleatoric map largely re-expresses the injected noise—the inverse crime [18]—so its near-perfect calibration is partly circular, and in contrast-to-noise terms it flags high-contrast structure rather than the low-contrast findings the Rose criterion marks as most vulnerable [4]. The epistemic map, untethered from the injected noise and only moderately correlated with the aleatoric one, is the honest vehicle for reconstruction uncertainty. Breaking the inverse crime—evaluating on a different degradation or real low-dose acquisitions—is the decisive next test.

Deployment profiles. Two configurations follow from the results. Arm S is the aleatoric-triage configuration: a single sub-50 ms forward pass returns the calibrated $\hat{\sigma}_a$ map at the full L2-level fidelity. Arm H adds the epistemic channel when needed, at the cost of $K=20$ stochastic passes (≈ 1 s); its epistemic map could additionally serve as a distribution-shift signal of the kind recent paediatric-AI reviews call for [15], though this is untested here. Translating the high-noise operating points (Arm A holds 32.9 dB at roughly $3\times$ reduced dose) into a dose-reduction claim requires a prospective reader study.

4.1 Limitations and future work

The clearest limitation is that the degradation is simulated in the image domain, so noise scales with displayed brightness and falls hardest on bone rather than on the photon-starved lung; a projection-domain model on raw data would be the single most improving change, and even then a faithful aleatoric map would relocate to the photon-starved anatomy, leaving low-contrast findings to the epistemic channel. The single-centre Kermany data is JPEG-compressed and demographically narrow, so generalisation to other collections (MIMIC-CXR [65], CheXpert) is unestablished; the 62.5/37.5 pneumonia/normal split inflates aggregate PSNR by ≈ 0.4 dB without changing rankings; labels are binary, precluding severity analysis; loss weights were not exhaustively searched; the three noise presets are fixed reference points rather than a sample of the clinical dose distribution; and the VAE findings rest on a single backbone, so the schedule-over-architecture conclusion is untested across other encoder–decoders. No pixel metric substitutes for a radiologist, and no reader study was performed.

Future work, in order of clinical impact: a blinded radiologist reader study on Arms A, D and H (diagnostic confidence, lesion detection, and whether the uncertainty overlay supports reader decisions); validation on real, full-resolution low-dose scans (e.g. MIMIC-CXR-JPG [65]); anatomy-attributed, thresholded per-scan uncertainty reports (Fig. 10 sketches this overlay for Arm H, with threshold calibration against confirmed pathology

the immediate next step); and, methodologically, β -TCVAE latent structure [66], a CXR-pretrained perceptual loss to replace the ineffective ImageNet VGG-16 [67], and deep ensembling [33] as an architecture-independent epistemic baseline.

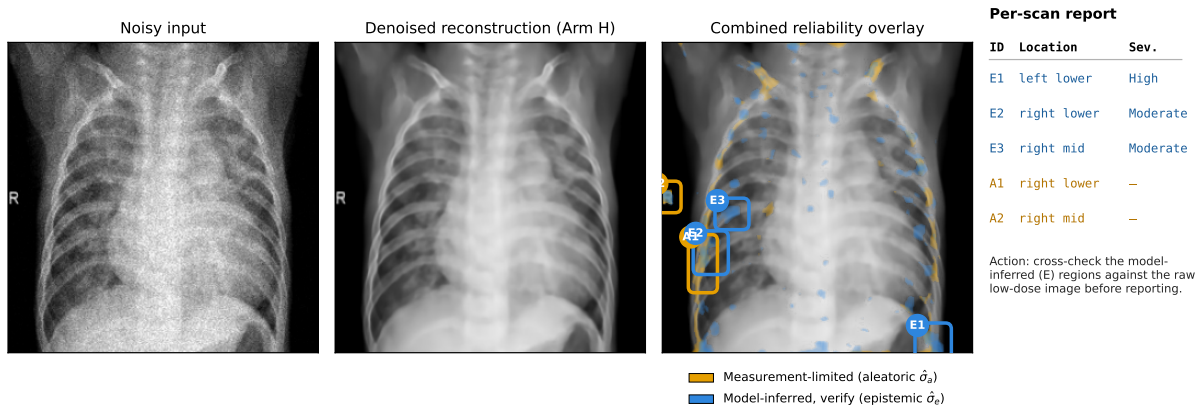


Figure 10: Prototype reliability-triage overlay (Arm H) on one representative case: noisy low-dose input, denoised reconstruction, and a combined overlay carrying both uncertainty channels with zone-attributed bounding regions and a per-scan report. Measurement-limited regions (amber, aleatoric) track the bony thorax; model-inferred regions (blue, epistemic, “verify”) mark where the latent prior filled the reconstruction in and are the clinically actionable signal. A design sketch, not a validated tool.

5 Conclusions

PP-VAE-Hformer inserts a variational bottleneck and a heteroscedastic dual head into a hybrid convolution–transformer denoiser and trains with a composite NLL+SSIM+FFL+KL objective. Across a 19-arm ablation and eight retrained baselines on Poisson–Gaussian-degraded paediatric chest radiographs, it shows that a calibrated per-pixel uncertainty map can be attached at effectively no cost to reconstruction quality: the heteroscedastic head’s 1.2 dB price is erased by two-stage fine-tuning (0.010 dB, statistically indistinguishable), the KL-annealing schedule rather than the bottleneck decides whether the variational variant reaches CNN parity, and the aleatoric map is calibrated to within 1% of unity, subject to the inverse-crime caveat that positions the epistemic map as the genuine reconstruction-uncertainty signal. For low-dose paediatric chest radiography, the confidence of a reconstruction can be made as reportable as the reconstruction itself; what stands between this and the clinic is validation—a blinded reader study and real low-dose data—not method.

Abbreviations

ARI: adjusted Rand index; CNN: convolutional neural network; CXR: chest radiograph; FFL: focal frequency loss; FSIM: feature similarity index; FT: fine-tuning; GELU: Gaussian error linear unit; KL: Kullback–Leibler (divergence); LPIPS: learned perceptual image patch similarity; MSE: mean squared error; MS-SSIM: multi-scale structural similarity;

NLL: negative log-likelihood; PReLU: parametric rectified linear unit; PSNR: peak signal-to-noise ratio; ROI: region of interest; SSIM: structural similarity index; VAE: variational autoencoder.

Declarations

Ethics approval and consent to participate

Not applicable. This study used only the publicly available, fully de-identified Kermany paediatric chest radiograph collection [48]; no new human data were collected.

Consent for publication

Not applicable.

Availability of data and materials

The dataset supporting the conclusions of this article is the publicly available Kermany paediatric chest X-ray collection [48] (<https://data.mendeley.com/datasets/rsbjbr9sj/2>). The full training, evaluation and analysis code supporting the conclusions of this article is available in the PPVAE-Hformer-codebase repository, <https://github.com/USN4730/PPVAE-Hformer-codebase>.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was carried out while STM was a Mastercard Foundation Scholar at the University of Cambridge, undertaking the MPhil in Population Health Sciences (specialising in Health Data Science) and based at St Edmund’s College. The Mastercard Foundation Scholarship supported STM during this research. The funder had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Authors’ contributions

STM and KF conceived the study. STM designed the study, implemented the models, performed the experiments and statistical analyses, and drafted the manuscript. KF supervised the work and revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

Model training used the Cambridge Service for Data Driven Discovery (CSD3).

References

- [1] John D Mathews, Anna V Forsythe, Zoe Brady, Martin W Butler, Scott K Goergen, Graham B Byrnes, Graham G Giles, Annette B Wallace, Peter R Anderson, Timothy A Guiver, Paul McGale, Timothy M Cain, James G Dowty, Adrian C Bickerstaffe, and Sarah C Darby. Cancer risk in 680,000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians. *BMJ*, 346:f2360, 2013. doi: 10.1136/bmj.f2360.
- [2] Michael Hauptmann, Graham Byrnes, Elisabeth Cardis, Marie-Odile Bernier, Maria Blettner, Jérémie Dabin, Hilde Engels, Jordi Figuerola, Tore Sanner Istad, Dušan Kopecý, Thomas Kröll, Johanna M Meulepas, Anna Nordenskjöld, Joachim Schüz, Isabelle Thierry-Chef, Thomas Tichler, Amy Berrington de González, and Mark P Little. Brain cancer after radiation exposure from CT examinations of children and young adults: results from the EPI-CT cohort study. *The Lancet Oncology*, 24(1): 45–53, 2023. doi: 10.1016/S1470-2045(22)00655-6.
- [3] Rebecca Smith-Bindman, Michael L. Kwan, Emily C. Marlow, Antonio Ortiz, Diana L. Miglioretti, et al. Medical imaging and pediatric and adolescent hematologic cancer risk. *New England Journal of Medicine*, 393:1269–1278, 2025. doi: 10.1056/NEJMoa2502098.
- [4] Albert Rose. The sensitivity performance of the human eye on an absolute scale. *Journal of the Optical Society of America*, 38(2):196–208, 1948. doi: 10.1364/JOSA.38.000196.
- [5] Jerrold T. Bushberg, J. Anthony Seibert, Edwin M. Jr. Leidholdt, and John M. Boone. *The Essential Physics of Medical Imaging*. Lippincott Williams & Wilkins, Philadelphia, 3rd edition, 2012.
- [6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. doi: 10.1109/TIP.2007.901238.
- [7] 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.
- [8] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2022. doi: 10.1109/TPAMI.2021.3088914.
- [9] Jingyun Liang, Jiezhong Cao, Guolu Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. doi: 10.1109/ICCVW54120.2021.00210.
- [10] Shi-Yu Zhang, Zhao-Xuan Wang, Hai-Bo Yang, Yi-Lun Chen, Yang Li, Quan Pan, Hong-Kai Wang, and Cheng-Xin Zhao. Hformer: Highly efficient vision transformer

- for low-dose CT denoising. *Nuclear Science and Techniques*, 34(4):61, 2023. ISSN 1001-8042. doi: 10.1007/s41365-023-01208-0.
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.
 - [12] Qi Gao, Zilong Li, Junping Zhang, Yi Zhang, and Hongming Shan. CoreDiff: Contextual error-modulated generalized diffusion model for low-dose CT denoising and generalization. *IEEE Transactions on Medical Imaging*, 43(2):745–759, 2024. doi: 10.1109/TMI.2023.3320812.
 - [13] Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. *IEEE Transactions on Medical Imaging*, 40(11):3169–3180, 2021. doi: 10.1109/TMI.2021.3077857.
 - [14] Laura Pfaff, Fabian Wagner, Nastassia Vysotskaya, Mareike Thies, Noah Maul, Siyuan Mei, Tobias Wuerfl, and Andreas Maier. No-New-Denoiser: A critical analysis of diffusion models for medical image denoising. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2024*, Lecture Notes in Computer Science. Springer, 2024. doi: 10.1007/978-3-031-72117-5_53.
 - [15] Lil B. Laborie, J. Lee, E. Antram, et al. Performance of adult-trained artificial intelligence models in paediatric imaging—a scoping review. *European Radiology*, 36: 5916–5939, 2026. doi: 10.1007/s00330-026-12354-5. Scoping review; 20 studies met inclusion criteria.
 - [16] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. doi: 10.1109/TCI.2016.2644865.
 - [17] Gustav Bredell, Kyriakos Flouris, Kerem Can Tezcan, and Ender Konukoglu. Explicitly minimizing the blur error of variational autoencoders. In *International Conference on Learning Representations*, 2023.
 - [18] Jari Kaipio and Erkki Somersalo. Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504, 2007. doi: 10.1016/j.cam.2005.09.027.
 - [19] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3929–3938, 2017. doi: 10.1109/CVPR.2017.300.
 - [20] 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.
 - [21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

- [22] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 17–33, 2022. doi: 10.1007/978-3-031-20071-7_2.
- [23] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhong Cao, Yulun Zhang, Hao Tang, Radu Timofte, and Luc Van Gool. Practical blind denoising via swin-conv-UNet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023. doi: 10.1007/s11633-023-1466-0.
- [24] Ilerioluwakiiye Abolade, Emmanuel Idoko, Solomon Odelola, Promise Omoigui, Adetola Adebawo, Aondana Iorumbur, Udunna Anazodo, Alessandro Crimi, and Raymond Confidence. SharpXR: Structure-aware denoising for pediatric chest X-rays. *arXiv preprint arXiv:2508.08518*, 2025.
- [25] Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. *Proceedings of IEEE International Conference on Image Processing*, 2:168–172, 1994. doi: 10.1109/ICIP.1994.413553.
- [26] 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.
- [27] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multi-scale structural similarity for image quality assessment. In *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216.
- [28] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 13919–13929, 2021. doi: 10.1109/ICCV48922.2021.01366.
- [29] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9023–9032, 2020. doi: 10.1109/CVPR42600.2020.00791.
- [30] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 5301–5310, 2019.
- [31] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [32] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1184–1193. PMLR, 2018.

- [33] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [34] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.
- [35] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 240–250, 2019. doi: 10.18653/v1/N19-1021.
- [36] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 10–21, 2016. doi: 10.18653/v1/K16-1002.
- [37] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Wierstra. Improving variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.
- [38] Pauline Shan Qing Yeoh, Khairunnisa Hasikin, Xiang Wu, and Siew Li Goh. Trends and applications of variational autoencoders in medical imaging analysis. *Computerized Medical Imaging and Graphics*, page 102647, 2025. doi: 10.1016/j.compmedimag.2025.102647.
- [39] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks, 2019. URL <https://arxiv.org/abs/1901.08644>. arXiv preprint arXiv:1901.08644.
- [40] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical Poissonian-Gaussian noise modelling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. doi: 10.1109/TIP.2008.2001399.
- [41] Sangyoon Lee, Min Seok Lee, and Moon Gi Kang. Poisson–Gaussian noise analysis and estimation for low-dose X-ray images in the NSCT domain. *Sensors*, 18(4):1019, 2018. doi: 10.3390/s18041019.
- [42] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. doi: 10.1007/978-3-030-01261-8_1.
- [43] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs), 2016. arXiv preprint arXiv:1606.08415.
- [44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.
- [46] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR)*, 2016. arXiv:1511.05440.
- [47] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.
- [48] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina C S Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fang Yan, Justin Dong, Made K Prasadha, Jacqueline Pei, Magdalene Y L Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Ruiling Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A N Huu, Cindy Wen, Edward D Zhang, Charlotte L Zhang, Oulan Li, Xiaobo Wang, Michael A Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M Anthony Lewis, Huimin Xia, and Kang Zhang. 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.
- [49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- [50] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.
- [51] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8): 2378–2386, 2011. doi: 10.1109/TIP.2011.2109730.
- [52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. doi: 10.1109/CVPR.2018.00068.
- [53] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022. doi: 10.3390/s22155540.
- [54] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.
- [55] B. L. Welch. The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947. doi: 10.1093/biomet/34.1-2.28.
- [56] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.

- [57] Gail M. Sullivan and Richard Feinn. Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, 4(3):279–282, 2012. doi: 10.4300/JGME-D-12-00156.1.
- [58] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2 edition, 1988.
- [59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [60] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17: 261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- [61] Takuya Akiba, Shotaro Sano, Toshihiro Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 2019. doi: 10.1145/3292500.3330701.
- [62] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6228–6237, 2018. doi: 10.1109/CVPR.2018.00652.
- [63] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.
- [64] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.
- [65] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0.
- [66] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL <https://arxiv.org/abs/1802.04942>.

- [67] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711, 2016. doi: 10.1007/978-3-319-46475-6_43.