Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sharp but Wrong? Fidelity and Hallucination in Deep CT Super-Resolution with Analytic Phantoms

Tanvi Sharma
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee

Research Article

Keywords: Deep super-resolution, X-ray computed tomography (CT), EDSR, ESRGAN, image reconstruction, hallucination, perceptual loss, PSNR, LPIPS, Nyquist limit, aliasing.

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11685421/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Sharp but Wrong? Fidelity and Hallucination in Deep CT Super-Resolution with Analytic Phantoms

Tanvi Sharma[1] and Mayank Goswami,[*1,2]

[1]Department of Physics, IIT Roorkee
[2]MF School of AI and DS, IIT Roorkee
mayank.goswami@ph.iitr.ac.in

Abstract

Deep super-resolution (SR) could recover resolution that X-ray CT trades for dose and scan time, but in CT, detail that looks plausible yet is not physically present is a failure, and standard metrics cannot see it. We train a pixel-loss residual network (EDSR) and an adversarially trained ESRGAN to upsample CT slices and evaluate both reconstruction fidelity and the risk of hallucinated detail. For the latter we use analytic phantoms with line-pair gratings and pores of exactly known frequency and diameter, held-out frequency bands, a documented blur–decimation–noise model, and a validated spectral measurement that labels each grating resolved, washed out, aliased or hallucinated. At $\times 4$, all networks improve on bicubic and Lanczos interpolation by more than 6 dB PSNR, and ESRGAN trades about 2 dB for the best perceptual score (LPIPS 0.015 vs. 0.046 for EDSR), yet it hallucinates 61% of the gratings the input cannot represent, where interpolation hallucinates none. An adversarial-weight sweep shows that the perceptual loss alone produces these inventions. Pixel-loss networks replace held-out gratings just below the Nyquist limit with their alias partner from training. At $\times 2$ hallucination falls sharply but does not vanish for ESRGAN (27%). Pores, and thin lines never seen in training, are recovered by every network. All data are synthetic and reproducible from a seed.

I. Introduction

Industrial and medical CT acquisitions trade spatial resolution against dose, scan time and field of view. Single-image super-resolution, which learns to map a coarse slice to a finer one from paired examples, promises to recover some of that resolution after the fact (Dong et al., 2014; Lim et al., 2017; Umehara et al., 2018; You et al., 2020). Generative adversarial training makes the output markedly sharper (Ledig et al., 2017; Wang et al., 2018), but an SR network can only output detail that is either present in the input or supplied by its learned prior. Where the input carries no information, as above the Nyquist limit of the LR grid (Shannon, 1949), any detail in the output is, by construction, an invention (Baker and Kanade, 2000).

For natural images such invention is often acceptable. For CT it is not: super-resolved slices are used to measure wall thicknesses, pore sizes and defect populations, and a convincingly rendered but fictitious pore or periodic texture can change a pass/fail decision. The difficulty is that the usual evaluation does not detect it. PSNR and SSIM (Wang et al., 2004) prefer conservative, blurred estimates; LPIPS (Zhang et al., 2018) and other perceptual measures prefer sharp ones, whether or not the sharpness is correct. This is the perception–distortion trade-off (Blau and Michaeli, 2018) viewed from the side of the application: neither axis measures whether a structure is *true*.

We make this measurable. Our protocol replaces visual judgement with analytic test objects whose ground truth is exact, and asks of every structure whether the output contains it, washes it out, aliases it, or replaces it with something else. Our contributions are:

- A reproducible data pipeline: fine-structure phantoms with square-wave gratings (1–8 lp/mm) and pores (2–10 px), rendered with $8\times$ supersampling, and a documented blur–decimation–Poisson–Gaussian degradation. Some frequency and diameter bands are withheld from training, and splits are made by phantom or volume, never by slice; a separate test set of thin isolated lines probes a structure absent from training (§III).
- A hallucination measurement that separates four outcomes (resolved, washed out, aliased by the LR sampling, and hallucinated), validated on the noise-free ground truth and with positive controls, and complemented by a measurement-domain consistency check (§IV).
- An empirical comparison of bicubic and Lanczos interpolation, EDSR (Lim et al., 2017), a PSNR-trained RRDB generator and ESRGAN (Wang et al., 2018), including an adversarial-weight sweep and a $\times 2$ vs. $\times 4$ comparison

*Corresponding author: mayank.goswami@ph.iitr.ac.in

(§VI).

On fidelity, all networks beat interpolation by more than 6 dB PSNR; the PSNR-trained RRDB generator is the most faithful, and ESRGAN gives up about 2 dB for the best LPIPS. On hallucination, the main findings are that the network with the best perceptual score is the one that most often invents gratings the input cannot contain, and that pixel-loss networks also hallucinate, though in a more specific way: when a held-out frequency has an alias partner that was seen in training, they output the partner.

II. Related Work

a) Single-image super-resolution.: Convolutional SR began with SRCNN (Dong et al., 2014); sub-pixel convolution (Shi et al., 2016) and deep residual designs such as EDSR (Lim et al., 2017) established strong pixel-loss baselines (see Wang et al., 2021b for a survey). SRGAN (Ledig et al., 2017) and ESRGAN (Wang et al., 2018) added adversarial and perceptual losses (Johnson et al., 2016; Simonyan and Zisserman, 2015), with ESRGAN introducing residual-in-residual dense blocks (RRDB) and a relativistic discriminator (Jolicoeur-Martineau, 2019). Real-ESRGAN (Wang et al., 2021a) emphasised the role of the degradation model used to synthesise training pairs, a point that applies equally to our setting.

b) Super-resolution for CT.: CNN-based SR has been applied to chest CT (Umehara et al., 2018), and GAN-based methods with cycle and residual constraints have been proposed for CT SR (You et al., 2020). These works are typically evaluated with PSNR, SSIM and reader studies, which do not separate correct from fabricated detail.

c) Hallucination in imaging.: Baker and Kanade (2000) used "hallucination" approvingly for learned face SR. In medical imaging the word now names a risk: distribution-matching losses can add or remove features in image translation (Cohen et al., 2018), deep reconstruction can be unstable to small perturbations (Antun et al., 2020), and Bhadra et al. (2021) define hallucinations in tomographic reconstruction through the null space of the imaging operator, that is, as error components that the measurements cannot detect. Our measurement-domain check (§IV-0c) adopts this view. Our phantoms make the question concrete and quantitative for SR, in the spirit of the line-pair targets used in CT quality assurance.

III. Data and Degradation Model

All data are synthetic and generated from a single seed; Figure 1 shows examples and Appendix A lists every parameter.

a) Fine-structure phantoms.: Each 512×512 slice is a rounded square of matrix material with linear attenuation $\mu = 0.075\,\mathrm{mm}^{-1}$ (aluminium at an effective energy of about 60 keV) on air, carrying a smooth Gaussian-random-field texture (standard deviation 3% of μ). A 4×4 grid of 96 px tiles holds either a square-wave grating or a 3×3 array of circular pores ($\mu = 0$). With a nominal pixel pitch of 0.05 mm, grating frequencies span 1–8 lp/mm (periods 20 to 2.5 px) with orientation $0°, 45°, 90°$ or $135°$ and contrast 0.4–1.0; pore diameters span 2–10 px. Everything is rendered at $8\times$ supersampling and area-averaged, so the ground truth is partial-volume correct. The generator records every element's frequency, orientation, phase, contrast, centre and diameter. At $\times4$ the LR pitch is 0.2 mm and the Nyquist limit is 2.5 lp/mm; because uniform sampling over 1–8 lp/mm would put most gratings above it, half of all gratings are drawn from 1–2.5 lp/mm and half from 2.5–8 lp/mm.

b) Held-out bands.: Frequencies in [2.0,2.5), [4.0,4.5) and [7.0,8.0] lp/mm and pore diameters in [3.5,4.5) and [6.5,7.5) px never occur in training or validation phantoms. Test phantoms sample them with probability 0.5, so the test set contains both trained and held-out structure, below and above the Nyquist limit.

c) Texture slices.: To train on irregular content and to report conventional fidelity metrics, we generate multi-phase porous textures (pores, matrix and dense inclusions obtained by thresholding Gaussian random fields, plus fine grain texture, inside a cylindrical sample). These are procedural stand-ins for real clean CT; we discuss this limitation in the Limitations section.

d) Degradation.: An HR slice $x \in [0,1]$ is degraded as

$$y = \tfrac{1}{N_0}\,\mathrm{Poisson}(N_0\,Ax) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2), \quad (1)$$

where A is a Gaussian blur ($\sigma_b \sim U[1.0, 1.5]$ px, 5×5 kernel) followed by area-averaging decimation by the SR factor, $N_0 = 10^5$ and $\sigma = 0.01$: a heteroscedastic Poisson–Gaussian model (Foi et al., 2008). Area averaging, rather than pixel skipping, mimics a detector that integrates over a larger pixel. On the built data the empirical LR noise variance matches the model to a ratio of 0.998.

e) Splits.: We generate 200 phantoms and 400 texture slices from 22 procedural volumes, split 70/15/15 by phantom or volume (140/30/30 phantoms; 286/57/57 texture slices). Fixed LR files serve validation and testing. For training, LR patches are re-simulated after each flip or rotation, so every epoch sees fresh, physically consistent noise.

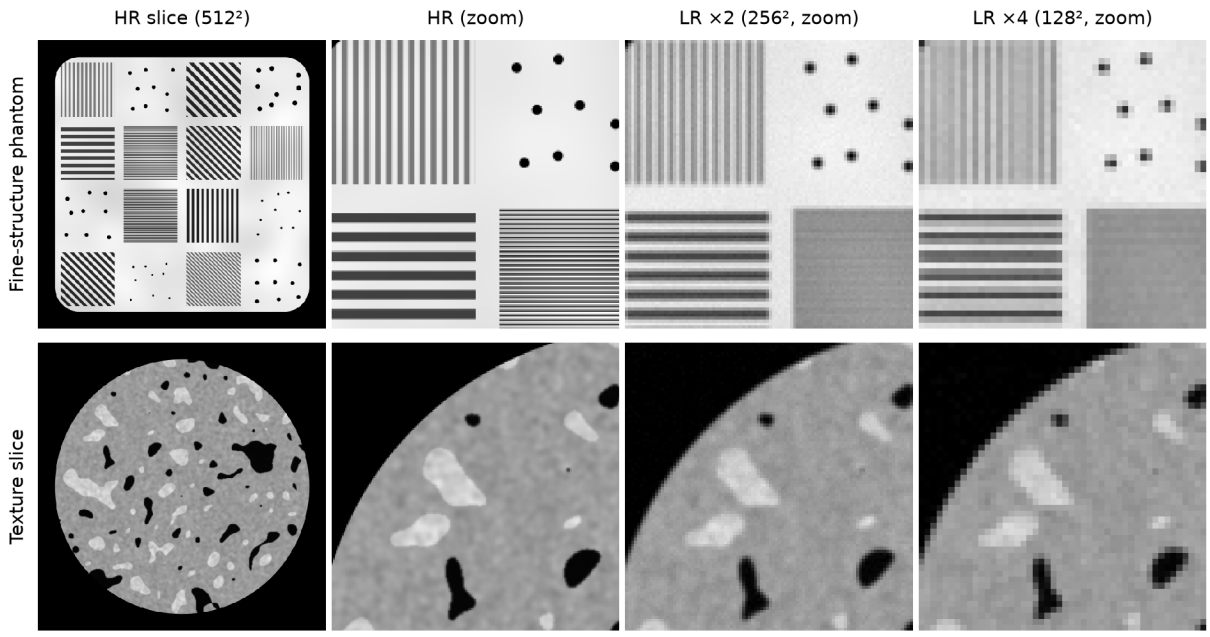


Fig. 1: Data. Top: a test fine-structure phantom (4×4 tiles of gratings and pore arrays in an aluminium-like matrix); bottom: a procedural texture slice. Columns: full HR slice, HR zoom, and the corresponding regions of the ×2 and ×4 LR inputs (nearest-neighbour display). Gratings above the LR Nyquist limit (bottom-right tile of the zoom) are washed out in the ×4 input.

f) Thin-line test set.: The specification also asks for thin lines, a structure our training phantoms do not contain. We therefore build a separate test set of 30 phantoms with the same matrix, texture, normalisation and degradation, in which every tile holds three parallel slits (crack-like lines) of one width $w \in [0.5, 6]$ px, contrast 0.4–1.0 and orientation, 24 px apart so that they are isolated lines rather than a grating. At ×4, lines with $w < 4$ px are narrower than one LR pixel. No model is retrained for this test.

IV. Measuring Hallucination

All measurements are taken on the super-resolved slice at HR size, tile by tile, against the exact phantom specification.

a) Gratings.: After trimming 4 px from each tile edge, we subtract the mean, apply a 2D Hann window, zero-pad to 1024^2 and locate the strongest spectral peak above 0.025 cycles/px with parabolic sub-bin refinement, giving a measured frequency and orientation. Two lock-in amplitudes complete the description: the *modulation ratio*, the amplitude at the true frequency relative to that of the ground truth, and the *wrong-pattern strength*, the amplitude at the measured frequency relative to the same reference. Each grating then receives one outcome:

- **resolved**: measured frequency within 10% of the truth and modulation ratio ≥ 0.2;
- **aliased**: a strong (≥ 0.2) pattern at a wrong frequency that coincides with where the LR sampling grid folds the grating or its 3rd or 5th square-wave harmonic, i.e. an error already present in the input;
- **hallucinated**: a strong pattern at a wrong frequency that the LR sampling does not explain;
- **washed out**: none of the above, i.e. no strong periodic structure.

Distinguishing aliasing from hallucination matters: interpolation of an aliased input faithfully reproduces the alias, which is a property of the acquisition rather than of the method.

b) Pores.: We threshold each pore tile at half its median (matrix) level, label 4-connected components, and match them to the true centres (tolerance $d/2 + 2$ px). This gives detected, missed and spurious counts and an equivalent diameter $\sqrt{4A/\pi}$. A blur-insensitive *integrated-deficit* diameter, obtained from the missing attenuation summed in a disc of radius $d/2 + 3$ px, separates "the pore has the wrong size" from "the pore is blurry".

c) Measurement-domain consistency.: Following Bhadra et al. (2021), we also ask whether an error is visible to the measurement. For each tile we compute the RMSE between the SR output and the ground truth both in the image domain and after applying the noise-free operator A of Eq. 1 to both. Content with large image-domain error but measurement-domain error below the noise level

lies (approximately) in the null space of A: the input cannot rule it out.

d) Thin lines.: Pixels of a line tile are binned by their coordinate perpendicular to the lines, giving a 1-D profile. At each true line we measure the dip depth below the background relative to the ground truth (a line counts as detected at $\geq$0.2, within $w/2 + 2$ px of its position) and its full width at half maximum. A *ghost line* is a dip away from all true lines whose prominence is at least 0.2 of the true line depth; on the ground truth and on interpolated outputs there are none.

e) Validation.: Applied to the noise-free ground truth, the grating measurement has a median frequency error of 0.01% (maximum 0.23%) and labels every grating resolved; the integrated pore diameter has a median error of 0.6%. Unit tests provide positive controls: a grating replaced by one at 1.3× the true frequency is labelled hallucinated, and one replaced by its alias frequency is labelled aliased.

V. Models and Training

a) Interpolation.: Bicubic and Lanczos (OpenCV) upsampling serve as non-learned baselines.

b) EDSR.: 16 residual blocks without batch normalisation (64 channels, residual scaling 0.1), a global skip connection and a pixel-shuffle upsampler (Lim et al., 2017; Shi et al., 2016); 1.5M parameters; L1 loss.

c) RRDB-PSNR and ESRGAN.: The ESRGAN generator (Wang et al., 2018) has 23 RRDBs (64 channels, growth 32; 16.9M parameters). It is first trained with L1 loss for 100 epochs (RRDB-PSNR, used both as a baseline and as the warm start), then fine-tuned for 80 epochs with

$$\mathcal{L} = \lambda_{\text{pix}}\,\mathcal{L}_1 + \lambda_{\text{perc}}\,\mathcal{L}_{\text{VGG}} + \lambda_{\text{adv}}\,\mathcal{L}_{\text{RaGAN}}, \quad (2)$$

with $\lambda_{\text{pix}} = 0.01$, $\lambda_{\text{perc}} = 1$ and $\lambda_{\text{adv}} = 0.005$ by default. $\mathcal{L}_{\text{VGG}}$ is the L1 distance between frozen ImageNet VGG19 conv5_4 features (pre-activation; grey values replicated to three channels) (Simonyan and Zisserman, 2015; Johnson et al., 2016), and $\mathcal{L}_{\text{RaGAN}}$ is the relativistic average adversarial loss (Jolicoeur-Martineau, 2019) with a 7-layer PatchGAN discriminator (Isola et al., 2017) that outputs an 8×8 logit map for a 128 px patch.

d) Training.: Each epoch draws 16 random 128×128 HR patches per training slice (6,816 patches; 32×32 LR at ×4), batch size 16, Adam (Kingma and Ba, 2015) with learning rate 10^{-4} halved every 40 epochs. EDSR trains for 200 epochs. Pixel-loss models are selected by validation PSNR on texture slices; GAN models are taken at the final epoch, since PSNR-based selection would undo the purpose of adversarial training. For ×2 we use 64 px HR patches so that the LR patch, and hence the cost per step, is unchanged. All models are implemented in PyTorch (Paszke et al., 2019) and trained on a single NVIDIA A10G GPU (about 0.5 h for EDSR, 2.4 h for the warm start and 2.4 h for each GAN fine-tuning).

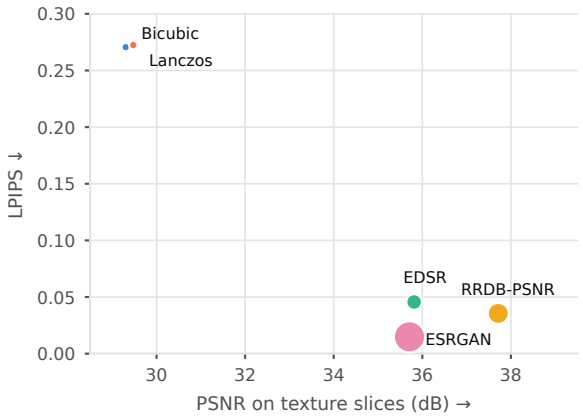


Fig. 2: Perception–distortion plane on test texture slices (×4). Marker area is proportional to the share of held-out-band gratings hallucinated.

VI. Results

A. Fidelity versus perception

Table I reproduces the familiar picture on texture slices. All networks improve on interpolation by more than 6 dB PSNR. RRDB-PSNR is the most faithful (37.71 dB, SSIM 0.953), and ESRGAN gives up ≈2 dB of it to reach by far the best LPIPS (0.015 vs. 0.036). Figure 2 places the methods in the perception–distortion plane (Blau and Michaeli, 2018). Judged by these metrics alone, ESRGAN would be the method of choice for a reader who values sharpness.

B. What the phantoms reveal

Figures 3 and 4 and Table II show what the texture metrics hide.

a) Above the Nyquist limit, ESRGAN invents.: Gratings above 2.5 lp/mm are attenuated by the blur and folded by decimation, so the LR input contains no usable trace of them. Interpolation and EDSR correctly return almost featureless tiles (93% and 79% washed out). ESRGAN instead fills 61% [55, 67] of these tiles with confident gratings, concentrated at 5–6 lp/mm regardless of the true frequency (Figure 4, right). RRDB-PSNR, its warm start, does so for 13%, so most of this behaviour is introduced by the fine-tuning; §VI-C shows that the perceptual loss alone suffices. A minority of above-Nyquist gratings is "resolved" by the networks (up to 36%). Because the input cannot determine these frequencies, such successes reflect

Method	Texture slices ($n = 57$)			Gratings on test phantoms ($n = 30$), %			Holes, %
	PSNR↑	SSIM↑	LPIPS↓	Halluc. held-out↓	Halluc. >Nyq.↓	Washed out >Nyq.	Found <1 LR px↑
Bicubic	29.30	0.845	0.271	0 [0,0]	0 [0,0]	93	0
Lanczos	29.47	0.840	0.272	0 [0,0]	0 [0,0]	90	0
EDSR	35.81	0.945	0.046	9 [5,15]	0 [0,0]	79	75
RRDB-PSNR	**37.71**	**0.953**	0.036	20 [15,26]	13 [8,19]	51	98
ESRGAN	35.71	0.929	**0.015**	49 [44,55]	61 [55,67]	4	100

TABLE I: Main results at ×4. Fidelity and perceptual metrics are means over test texture slices. Grating columns give the share of grating tiles labelled *hallucinated* (a strong pattern at a frequency explained by nothing in the input), with 95% cluster-bootstrap intervals over test phantoms; "held-out" gratings lie in frequency bands absent from training, ">Nyq." gratings lie above the LR Nyquist limit (2.5 lp/mm) and are therefore not recoverable from the input. The last column is the share of holes narrower than one LR pixel that are detected. Best value per column in bold (where a direction is defined).

Method	Trained (102)		Held-out (179)		<Nyq. (135)		>Nyq. (146)	
	Res.	Hal.	Res.	Hal.	Res.	Hal.	Res.	Hal.
Bicubic	50	0	47	0	96	0	3	0
Lanczos	53	0	47	0	97	0	5	0
EDSR	66	0	45	9	87	13	21	0
RRDB-PSNR	75	3	50	20	85	15	36	13
ESRGAN	77	17	51	49	88	12	35	61

TABLE II: Grating outcomes (%) at ×4 by frequency group (number of grating tiles in parentheses). Res. = resolved (true frequency within 10% and ≥20% of the true contrast); Hal. = hallucinated. The remainder is washed out or aliased.

Method	n	Alias partner (%)	Hallucinated: RMSE image	Hallucinated: RMSE meas.	Resolved meas. RMSE
EDSR	17	100	0.322	0.013	0.008
RRDB-PSNR	39	51	0.337	0.009	0.007
ESRGAN	105	21	0.353	0.039	0.018

TABLE III: Where hallucinated gratings come from (×4). *Alias partner*: share whose measured frequency gives the same LR samples as the true grating. Median per-tile RMSE against ground truth in the image domain and in the measurement domain (after re-applying the noise-free blur and decimation A). For reference, the LR noise standard deviation is ≈0.010.

a training prior that happens to match the truth, not recovered information.

b) Below the Nyquist limit, pixel-loss networks pick the wrong alias.: All networks resolve gratings at trained frequencies below the limit more sharply than interpolation (modulation ratio ≈1 vs. ≈0.5). The exception is the held-out band 2.0–2.5 lp/mm, just below the limit, where bicubic interpolation resolves 94% of the gratings while EDSR and RRDB-PSNR hallucinate 19% and 22% of them, outputting full-contrast gratings near $5 - f$ lp/mm (Figure 4, red points around 2.7 lp/mm). That frequency is f's alias partner on the ×4 grid: f and $5 - f$ lp/mm produce the same LR samples, and only the partner (2.5–3.0 lp/mm) occurred in training. Faced with an input that is ambiguous by construction, the networks resolve the ambiguity in favour of the training distribution. Table III quantifies this: 100% of EDSR's and 51% of RRDB-PSNR's hallucinated gratings are alias partners of the truth, against 21% for ESRGAN.

c) Can the input rule these hallucinations out?: Re-applying the noise-free degradation operator A separates two kinds of hallucination (Table III). Hallucinated tiles are far from the ground truth in the image domain for every network (median RMSE 0.322–0.353). For the pixel-loss networks they nevertheless lie as close to it in the measurement domain (EDSR 0.013, RRDB-PSNR 0.009) as their correctly resolved tiles do, and close to the LR noise level (≈0.010). In the terminology of Bhadra et al. (2021) these are null-space hallucinations: the alias partner reproduces the LR samples, so no data-consistency check on the LR image can detect them. ESRGAN's hallucinations are different: their measurement-domain error (0.039) is about four times the noise level and twice that of its own resolved tiles (0.018), so re-degrading the output and comparing with the input would flag many of them. A consistency check is therefore a useful but incomplete guard. It catches inventions that contradict the input, not the ones that the input cannot distinguish from the truth.

d) Pores are robust.: Table I (last column) and Figure 3(d) show the reverse case. Isolated pores narrower than one LR pixel are lost by interpolation (0% detected) but recovered by every network (EDSR 75%, RRDB-PSNR 98%, ESRGAN 100%), with no more spurious detections than the ground truth itself and a median threshold-diameter error of 0.8% (RRDB-PSNR) and 0.7% (ESRGAN) for pores wider than an LR pixel. Unlike a dense grating, an isolated pore leaves a localised, well-conditioned trace in the LR image (its integrated attenuation deficit), so the problem is well posed and the learned prior helps rather than misleads.

e) Thin lines are recovered, but strong adversarial training adds ghosts.: Isolated thin lines, which no model saw in training, behave like

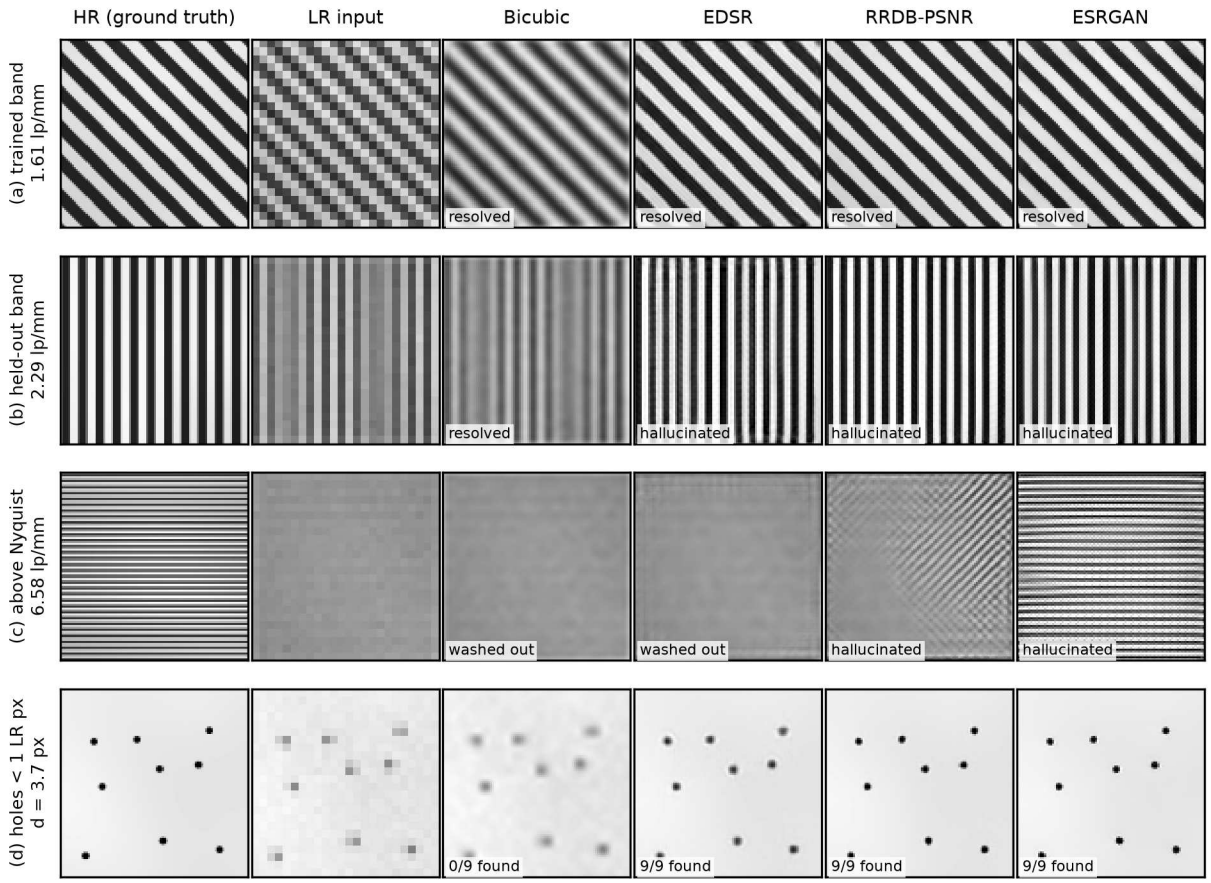


Fig. 3: One test tile per regime (×4), chosen by rule rather than by eye (first tile, in phantom order, that meets the regime's criterion). (a) Trained frequency below the LR Nyquist limit: all methods resolve it. (b) Held-out band just below the limit: interpolation resolves it, while all three networks output a sharp grating at a different frequency. (c) Above the limit, so the input carries no information: interpolation and EDSR return a flat tile, RRDB-PSNR a moiré, and ESRGAN a crisp but wrong grating. (d) Pores narrower than one LR pixel: every network recovers all nine. Labels give each method's measured outcome.

	×4, w <1 LR px			×4, w ≥1 LR px		×2, all	
Method	Depth	Δw	Ghost	Depth	Ghost	Depth	Ghost
Bicubic	0.44	+2.6	0	0.84	0	0.79	0
Lanczos	0.44	+2.5	0	0.84	0	0.79	0
EDSR	0.63	+1.1	0	0.97	0	0.96	0
RRDB-PSNR	0.62	+1.7	0	0.91	0	0.98	0
ESRGAN (λ_{adv}=0)	0.68	+1.8	1	0.96	0	–	–
ESRGAN	0.68	+1.6	3	0.95	0	0.99	0
ESRGAN (λ_{adv}=0.05)	0.72	+1.8	21	0.96	2	–	–

TABLE IV: Thin isolated lines, a structure never seen in training (30 test phantoms, 1,440 lines). *Depth*: median line depth relative to ground truth. Δw: median FWHM width error in px for lines with $w \geq 2$ px (positive: too wide). *Ghost*: % of tiles containing at least one line that is not there. Every method detects at least 98% of the lines.

pores rather than like gratings (Table IV). Every method detects at least 98% of them. For lines narrower than an LR pixel at ×4 the networks restore more of the true depth than interpolation (median 0.63 for EDSR and 0.68 for ESRGAN, against 0.44 for bicubic) and render them narrower, though still too wide (median width error +1.1 px for EDSR, +2.6 px for bicubic). Invented lines are rare: none for interpolation or the pixel-loss networks, and 3% of narrow-line tiles for ESRGAN at its default weight. The exception is the largest adversarial weight, $\lambda_{\text{adv}} = 0.05$, which adds ghost lines next to 21% of the narrow-line tiles. Unlike grating hallucination (§VI-C), this failure does depend on λ_{adv}. At ×2 no method produces ghost lines.

C. Adversarial weight

Table V (plotted in Appendix A, Figure 5) varies λ_{adv} over two orders of magnitude while keeping every other setting fixed. The perceptual score barely moves (LPIPS 0.016–0.019), and neither does hallucination: between 46% and 54% of held-out-band gratings and between 58% and 69% of above-Nyquist gratings are hallucinated, with overlapping confidence intervals and no monotone trend. Crucially, the run with $\lambda_{\text{adv}} = 0$, which keeps the VGG perceptual loss but has no discriminator,

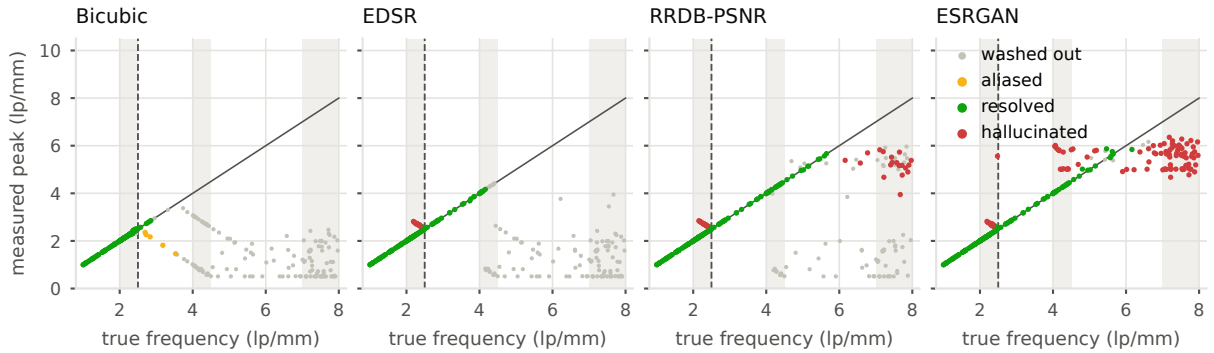


Fig. 4: Measured versus true grating frequency on test phantoms (×4), coloured by outcome. Shaded columns: held-out bands; dashed line: LR Nyquist limit (2.5 lp/mm); diagonal: perfect recovery. Washed-out tiles have no meaningful peak and are drawn small.

λ_{adv}	PSNR	SSIM	LPIPS	Hal. held-out	Hal. >Nyq.
warm start	37.71	0.953	0.036	20 [15,26]	13 [8,18]
0	35.77	0.831	0.019	46 [39,54]	66 [58,73]
0.001	36.08	0.844	0.017	51 [44,58]	63 [56,70]
0.005	35.71	0.929	0.015	49 [43,55]	61 [55,68]
0.01	30.18	0.935	0.017	50 [43,56]	58 [52,65]
0.05	34.29	0.866	0.016	54 [48,61]	69 [63,75]

TABLE V: Adversarial-weight sweep at ×4. All rows fine-tune the same PSNR-oriented warm start for 80 epochs with $\lambda_{pix} = 0.01$ and $\lambda_{perc} = 1$; $\lambda_{adv} = 0$ keeps the perceptual loss but removes the adversarial term. Hallucination in %, with 95% cluster-bootstrap intervals.

	×2 (Nyq. 5 lp/mm)				×4 (Nyq. 2.5 lp/mm)			
Method	PSNR	LPIPS	H-ho	H->N	PSNR	LPIPS	H-ho	H->N
Bicubic	32.41	0.177	0	0	29.30	0.271	0	0
EDSR	41.01	0.017	1	0	35.81	0.046	9	0
RRDB-PSNR	41.88	0.015	2	4	37.71	0.036	20	13
ESRGAN	39.89	0.006	13	27	35.71	0.015	49	61

TABLE VI: Upsampling factor. H-ho / H->N: % of held-out-band / above-LR-Nyquist gratings hallucinated. The Nyquist limit moves with the factor, so the ">Nyq." set differs between columns.

already hallucinates 66% of above-Nyquist gratings, against 13% for its warm start. In our setting the invented gratings are therefore a consequence of the perceptual objective (a feature-space loss with a pixel weight of only 0.01), not of the adversarial term, and lowering λ_{adv} is not a remedy. Fidelity varies non-monotonically across the sweep (PSNR 30.18–36.08 dB, SSIM 0.83–0.94). The low PSNR at $\lambda_{\text{adv}} = 0.01$ is partly a global intensity offset (about −0.017 on the slices we inspected) rather than structural error, which illustrates that single fine-tuning runs of this kind differ in ways that PSNR and SSIM conflate; we trained one seed per setting.

D. Upsampling factor

At ×2 the LR Nyquist limit doubles to 5 lp/mm, so two of the three held-out bands (2.0–2.5 and 4.0–4.5 lp/mm) become representable and only gratings from 5 lp/mm upward are lost by the sampling. Table VI shows that every method gains in PSNR and LPIPS, with the same ranking as at ×4, and that hallucination falls sharply. The pixel-loss networks almost stop hallucinating (EDSR 1%, RRDB-PSNR 2% of held-out-band gratings, against 9% and 20% at ×4): with the held-out bands now well inside the passband, the input determines the frequency, and the alias-partner substitution of §VI largely disappears. ESRGAN still hallucinates 13% of held-out-band and 27% of above-Nyquist gratings, against 0% for bicubic interpolation. The qualitative pattern is therefore unchanged: the failure tracks the information content of the input rather than the factor itself, and the perceptually trained model remains the only one that invents structure the input cannot contain. A milder factor shrinks the region where this can happen but does not remove it.

VII. Discussion

a) Metrics that cannot see the failure.: On texture slices the ranking by PSNR and the ranking by LPIPS disagree, which is expected, but neither is informative about whether structure is correct. The method with the best LPIPS is the one that most often invents structure, and it does so precisely where the input is uninformative. A procedure that selects SR models by LPIPS, or by visual preference, selects for this behaviour.

b) A practical audit.: The phantom protocol costs minutes per model and needs no annotation. We suggest reporting, alongside PSNR, SSIM and LPIPS, hallucination rates stratified by (i) whether the structure is representable at the LR Nyquist limit and (ii) whether its parameters occurred in training. The first stratum separates information recovery from prior-driven synthesis; the second separates generalisation from memorisation. A finding as specific as the alias-partner substitution is

invisible without both.

c) What SR can and cannot do in CT.: The results support a narrow but useful role for SR in quantitative CT: localised features with a well-conditioned LR trace, such as isolated pores, are recovered accurately and better than by interpolation. Dense periodic structure near or above the LR Nyquist limit should not be measured on super-resolved images, and models fine-tuned with perceptual or adversarial losses should not be used where fabricated texture would change a decision. Our sweep suggests that turning down the adversarial weight does not make such a model safe, because the perceptual loss alone produces the same inventions.

VIII. Conclusion

We introduced a phantom-based protocol that turns "does it hallucinate?" into a set of frequency-resolved, validated measurements, and used it, alongside PSNR, SSIM and LPIPS, to compare interpolation, EDSR, an RRDB generator and ESRGAN for CT super-resolution. By the standard metrics every network is far more faithful than interpolation, but those metrics do not reveal which networks invent structure. The perceptually best model is the least trustworthy above the Nyquist limit, a property that it inherits from the perceptual loss rather than from the adversarial term, pixel-loss models substitute alias partners seen in training for held-out frequencies, and isolated pores and thin lines are recovered by all networks, although a large adversarial weight adds ghost lines. Code, configuration and seeds suffice to regenerate every number in this paper.

Limitations

Synthetic data only. The texture slices are procedural stand-ins for real clean CT, drawn from only 22 volumes (three in the test split). Real anatomy or materials, reconstruction artefacts and scanner-specific noise texture are absent, so absolute PSNR, SSIM and LPIPS values, and the 6 dB gain over interpolation in particular, should not be expected to transfer. The phantom-based hallucination analysis does not depend on the texture source, but whether the specific failure modes appear with the same frequency on real data is an open question; open datasets such as the CWI walnut collection (Der Sarkissian et al., 2019) or the low-dose CT challenge data (McCollough et al., 2017) would allow this test. **Matched, known degradation.** Training and test LR images follow the same degradation model, in the image domain. Real resolution loss arises in the projection domain and through reconstruction; a mismatched or blind setting (Wang et al., 2021a) may change the results. **Simple phantoms.** Gratings, circular pores and straight lines are deliberately idealised. They make ground truth exact, but they cover only three structure types (the thin lines only at test time), and the alias-partner effect depends on the frequency bands we chose to withhold. **Scale.** Each configuration is trained once with one seed, and the test set has 30 phantoms (281 grating tiles, 199 pore tiles). We report cluster-bootstrap intervals over phantoms, but not variability across training runs. **Thresholds.** The outcome labels depend on thresholds (10% frequency tolerance, 0.2 modulation and strength). They were set while validating the measurement on ground truth and on the interpolation baselines, before any network was evaluated, and are checked by positive controls; other choices would shift the percentages.

Ethics Statement

This work uses no human or patient data. Its purpose is cautionary: it shows that super-resolution models, especially adversarially trained ones, can fabricate convincing structure that standard image-quality metrics reward. We discourage clinical or safety-critical use of such models without structure-level validation of the kind proposed here.

Video

https://youtu.be/gjZLoGGpHfs

References

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. 2020. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095.

Simon Baker and Takeo Kanade. 2000. Hallucinating faces. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 83–88.

Sayantan Bhadra, Varun A. Kelkar, Frank J. Brooks, and Mark A. Anastasio. 2021. On hallucinations in tomographic image reconstruction. *IEEE Transactions on Medical Imaging*, 40(11):3249–3260.

Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6228–6237.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. 2018. Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 529–536.

Henri Der Sarkissian, Felix Lucka, Maureen van Eijnatten, Giulia Colacicco, Sophia Bethany Coban,

and K. Joost Batenburg. 2019. A cone-beam X-ray computed tomography data collection designed for machine learning. *Scientific Data*, 6:215.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 184–199.

Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. 2008. Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711.

Alexia Jolicoeur-Martineau. 2019. The relativistic discriminator: A key element missing from standard GAN. In *International Conference on Learning Representations (ICLR)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 136–144.

Cynthia H. McCollough, Adam C. Bartley, Rickey E. Carter, et al. 2017. Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 low dose CT grand challenge. *Medical Physics*, 44(10):e339–e352.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035.

Claude E. Shannon. 1949. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

Kensuke Umehara, Junko Ota, and Takayuki Ishida. 2018. Application of super-resolution convolutional neural network for enhancing image resolution in chest CT. *Journal of Digital Imaging*, 31(4):441–450.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021a. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1905–1914.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision Workshops (ECCVW)*.

Zhihao Wang, Jian Chen, and Steven C. H. Hoi. 2021b. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3365–3387.

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

Chenyu You, Guang Li, Yi Zhang, et al. 2020. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE). *IEEE Transactions on Medical Imaging*, 39(1):188–203.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595.

Appendix

Grating tiles are analysed after a 4 px inset (88×88 px). The spectral search excludes frequencies below 0.025 cycles/px, which removes the DC term, Hann-window leakage and the slowly varying matrix texture. For the aliasing test, a frequency vector $\mathbf{f}$ is folded onto the LR grid as

Phantoms	
Slice size / pixel pitch	512^2 / 0.05 mm (nominal)
Supersampling	8×8
Tiles	4×4 of 96 px, gap 16 px
P(grating)	0.6
Grating frequency	1–8 lp/mm, half below 2.5
Orientation / contrast	0/45/90/135° / 0.4–1.0
Pore diameter	2–10 px, 9 per tile
Matrix μ / texture	0.075 mm^{-1} / s.d. 3%, corr. 24 px
Held-out frequencies	[2,2.5), [4,4.5), [7,8] lp/mm
Held-out diameters	[3.5,4.5), [6.5,7.5) px
Degradation	
Blur	Gaussian, $\sigma_b \sim U[1, 1.5]$, 5×5
Decimation	area average, ×4 and ×2
Noise	Poisson ($N_0 = 10^5$) + Gaussian ($\sigma = 0.01$)
Training	
Patches	16/slice/epoch, 128^2 HR (64^2 at ×2)
Optimiser	Adam, $\beta = (0.9, 0.99)$, lr 10^{-4}
Schedule	halve every 40 epochs
Epochs	EDSR 200; RRDB 100; GAN 80
Losses (GAN)	λ_{pix}=0.01, λ_{perc}=1, λ_{adv}=0.005
Seed	15 (data), 0 (training)

TABLE VII: Data generation, degradation and training settings.

$\mathbf{f} - \mathrm{round}(\mathbf{f}/f_s)\, f_s$ with $f_s = 1/s$ cycles per HR pixel for factor s, and compared with the measured peak up to sign; the 3rd and 5th harmonics of the square wave are folded in the same way. The alias-partner analysis of Table III applies the same folding to the measured peak and asks whether it lands on the fold of the true frequency (tolerance 0.006 cycles/px). Confidence intervals are 95% percentile intervals from 2,000 cluster-bootstrap resamples of test phantoms. Pore detection uses 4-connectivity, and the ground truth itself yields 0.07 spurious components per tile, which sets the floor for that metric.

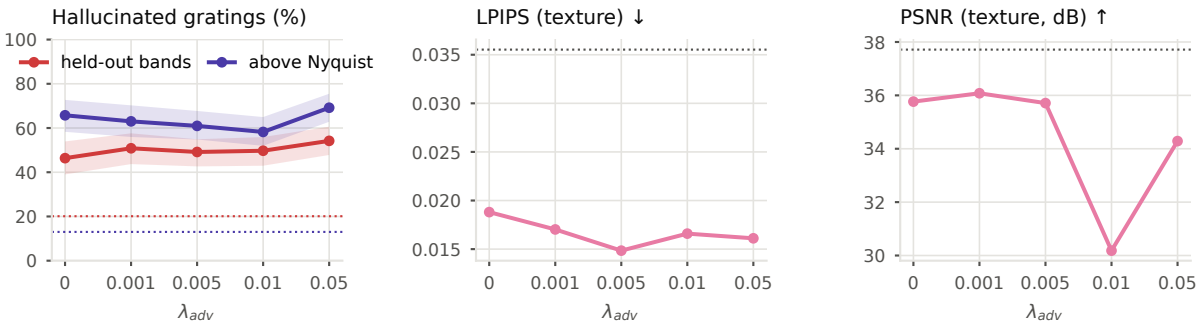


Fig. 5: Effect of the adversarial weight λ_{adv} at $\times 4$ (bands: 95% cluster-bootstrap intervals). Dotted lines: the RRDB-PSNR warm start that every run is fine-tuned from.