WHAT DOES HIGH-RESOLUTION SUPERVISION ACTUALLY **TEACH**? SEPARATING SOURCE SUPPORT FROM RESOLUTION SPECIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-resolution super-resolution corpora are motivated by the assumption that natively high-resolution images supply supervision that lower-resolution data cannot. Existing comparisons cannot test it: replacing a low-resolution corpus with a native-4K one changes the degradation operator, the source pixels available for sampling, the number of scenes, the field of view of a fixed crop, the allocation of compute and the train–test match all at once, so the reported difference is a sum of six effects and identifies none of them. We measure the raster's own contribution using same-source resolution pyramids, one declared degradation, a fixed optimisation budget, and evaluation of every arm both on a common test tier and across tiers. Source support, scene count and per-scene coverage are bound by one constraint, so no design varies the raster alone; we run all three that release one of them. On a source-rich corpus the apparent 4K advantage — $0.835\,\mathrm{dB}$ at 43M parameters — survives none of them, at any capacity or scale factor. On two unrelated RAW-derived corpora the preference between training rasters instead reverses with the target resolution. High-resolution data therefore supplies either additional source support or closer alignment with the deployment resolution, and the nominal raster measures neither.

1 INTRODUCTION

Datasets of natively high-resolution photographs have become a standard ingredient in super-resolution research, on intuitive reasoning. A 4K photograph records detail that a 1K photograph of the same scene cannot; a model trained only on the latter has never been shown that detail and so — the argument goes — cannot learn to restore it. On this view the raster of the training corpus is a proxy for the quality of its supervision, and acquiring at the highest available resolution is a dominant choice. It is also the expensive one, which makes it worth knowing whether the premise is true.

The premise has not been isolated. Comparisons that motivate high-resolution corpora replace one dataset with another, and a dataset is not a single variable: the swap also moves the degradation operator, the source pixels available to sample patches from, the number of distinct scenes, the fraction of a scene a fixed-size crop covers, how compute is distributed over the data, and the relationship between training and test image statistics. The reported difference is a sum over all six, and attributing it to resolution requires the other five to have contributed nothing — an assumption rarely stated, let alone checked. It does not hold in practice: auditing one published benchmark that releases its low-resolution images, we find them reproduced bit-exactly by a resampling operator different from the classical convention, an input mismatch worth $5.98\,\mathrm{dB}$ against the $4.68\,\mathrm{dB}$ the benchmark reports after adopting native-4K data — a factor the comparison did not hold fixed, and larger than the gain it reports.

This paper measures the quantity directly. We build same-source resolution pyramids in which every tier is one resampling of the same original, fix one degradation operator and optimisation budget across arms, intervene explicitly on the source-pixel budget that a raster change otherwise moves silently, and score every arm both on a common test tier and across tiers. The last matters as much as the first: scoring each arm on its own tier, the default, does not measure a contrast at all.

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

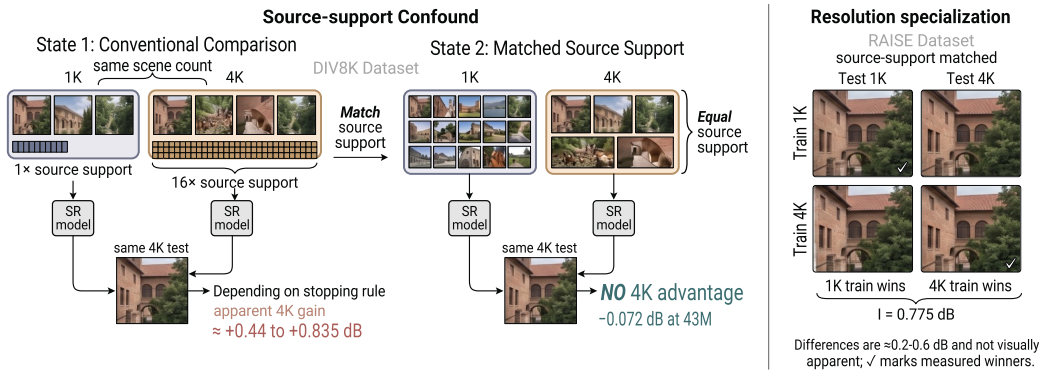

Figure 1: The same "4K advantage" has two sources and only one is about resolution. **Left:** the conventional comparison hands the native-4K corpus sixteen times the source pixels along with its raster and reports the sum; equalising that budget leaves $-0.072\,\mathrm{dB}$ at 43M, and no capacity or scale factor recovers an advantage. **Right:** on a RAW-derived corpus the same matched measurement does not vanish but *reverses*, each training raster winning at its own test tier.

Two findings follow, and the second bounds the first (Figure 1). On a source-rich corpus the apparent advantage of native-4K supervision grows steeply with model capacity, from $0.033\,\mathrm{dB}$ to $0.835\,\mathrm{dB}$, which reads as compelling evidence that larger models exploit a 4K-specific prior; matching the source-pixel budget removes it at every capacity and at a second scale factor. Matching cannot be complete — raster cannot move with source support, scene count and coverage all held, so some resource must always be released — so we run each of the three designs that releases one, and the advantage returns under none: at the capacity where the uncontrolled gap is largest, holding both the budget and the scene count leaves $-0.006\,\mathrm{dB}$. Yet on two unrelated RAW-derived corpora the preferred training raster reverses with the target resolution instead, over an interaction span of $0.775\,\mathrm{dB}$ to $0.807\,\mathrm{dB}$ that survives scene-grouped splitting, holds at two capacities, and does not move when scene counts are equalised. High-resolution supervision is therefore not intrinsically valuable — and not universally interchangeable with lower-resolution supervision either.

We make three contributions. **Identification:** we formalise why the conventional comparison is confounded, give a protocol that separates nominal raster from source support and from train–test alignment, show that the interaction span is invariant to the nuisance that otherwise makes datasets incomparable, and show that raster cannot be varied with source support, scene count and coverage all held, so the design space admits exactly three single-release configurations. **Measurement:** we run all three and find that the apparent 4K advantage in the source-rich regime survives none of them at any capacity or scale factor, while sources differ widely in resolution dependence. **Negative results:** six candidate predictors of where a corpus falls on that scale all fail, one of them against an acceptance interval recorded before measuring, which bounds what an eventual answer can look like.

2 Related Work

Ultra-high-resolution corpora, and the closest prior result. Progress in single-image super-resolution has come from architecture (Dong et al., 2014; Lim et al., 2017; Zhang et al., 2018b; Liang et al., 2021; Chen et al., 2023) and, more quietly, from a climb in the raster of the training corpus (Agustsson & Timofte, 2017; Gu et al., 2019b; Zhang et al., 2021b; Li et al., 2023). The evidence is typically a dataset swap, this paper's object rather than a result we dispute, and real-capture benchmarks (Cai et al., 2019; Wei et al., 2020) face the same difficulty from the other side; we therefore treat DIV8K, RAISE (Dang-Nguyen et al., 2015) and FiveK (Bychkovsky et al., 2011) as sources rather than benchmarks. Closest to us, Ohtani et al. (2024) find a curated low-resolution training set beats DF2K and LSDIR, the first demonstration we know of that competitive super-resolution needs no high-resolution images. It agrees with our source-rich result without identifying the cause, since DiverSeg differs from DF2K in raster, scene count, content and compression history at once.

Degradation and evaluation protocol. That the low-resolution operator dominates measured performance is not news: kernel mismatch outweighed the image prior a decade ago (Efrat et al., 2013), degradation pipelines exist because the bicubic convention does not transfer (Gu et al., 2019a; Bell-Kligler et al., 2019; Zhang et al., 2021a; Wang et al., 2021; Lugmayr et al., 2020), and an undeclared resampling detail can move generative metrics by more than the methods they adjudicate (Parmar et al., 2022; Zhang, 2019). We fix one declared operator across every tier, but it is not the confound we are after: held fixed, the raster still moves source support, scene count, crop field of view and train–test alignment, as do metric choice (Blau & Michaeli, 2018) and augmentation (Yoo et al., 2020).

Data scale, source support, and controlled comparison. The quantity we intervene on has a long history under another name: internal-statistics work established that a single natural image already supplies a large, redundant patch dictionary (Glasner et al., 2009; Zontak & Irani, 2011; Shocher et al., 2018), so a 4K photograph is sixteen times the sampling substrate. Scaling studies (Sun et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022) vary quantity at fixed composition and pruning shows composition can outweigh size (Sorscher et al., 2022); the resolution question is of that kind. Our estimator is difference-in-differences (Angrist & Pischke, 2009) on covariate shift (Shimodaira, 2000), and the paper belongs with controlled re-examinations that fix budget and protocol (Timofte et al., 2016; Henderson et al., 2018; Ferrari Dacrema et al., 2019; Musgrave et al., 2020).

3 What a Resolution Comparison Measures

Corpora, pipelines, and predictors. Write a training corpus as $\mathcal{D} = (\mathcal{X}, R, N)$: a distribution over scenes, the *nominal raster* at which each is stored, and the number of independent scenes drawn. Two derived quantities are routinely conflated with R: the *source support* $S(\mathcal{D}) = N\,|R|$, the pixels available to sample from, and the *exposure* $E = T \cdot B \cdot hw$, the pixels the optimiser consumes over T steps of batch B at crop $h \times w$. Under a fixed step budget these come apart completely — E is identical across arms by construction, while S scales with $|R|$. Training also fixes a degradation operator $\mathcal{K}$, a crop policy $\mathcal{C}$ and an optimisation budget $\mathcal{B}$; a learner returns $f = \mathcal{A}(\mathcal{D}, \mathcal{K}, \mathcal{C}, \mathcal{B})$, scored by $Q(f, t)$ on test condition t.

Why the usual comparison is not identified. Swapping a low-resolution dataset for a native-4K one moves an entire configuration $c = (\mathcal{X}, R, N, \mathcal{K}, \mathcal{C}, \mathcal{B})$ at once. Fix any ordering of the coordinates and interpolate through intermediates $c^{(0)} = c_a, \ldots, c^{(K)} = c_b$ differing in one coordinate at a time; the observed difference telescopes exactly,

$$\underbrace{Q(c_b) - Q(c_a)}_{\text{what a dataset swap reports}} \;=\; \sum_{k=1}^{K} \underbrace{\Big[\, Q\big(c^{(k)}\big) - Q\big(c^{(k-1)}\big) \Big]}_{\text{one coordinate moved}}, \tag{1}$$

so the reported number is a *sum* of single-factor effects, only one of which is the native-resolution effect. What the identity does not supply is uniqueness, and the size of that failure can be written down exactly.

Proposition 1 (Single-factor attribution is ambiguous by the interaction)**.** *Write Δ_i for the operator advancing coordinate i from its value in c_a to its value in c_b. Of the two orders in which coordinates i and j may be moved, one assigns to i the increment $\Delta_i Q\big|_{j=a}$ and the other $\Delta_i Q\big|_{j=b}$, and*

$$\Delta_i Q\big|_{j=b} \;-\; \Delta_i Q\big|_{j=a} \;=\; \Delta_j \Delta_i\, Q. \tag{2}$$

The term attributed to any one coordinate is therefore determined only up to the mixed second differences with the coordinates moved before it, and is well defined precisely when those vanish. (Both sides of equation 2 equal $Q(c^{ij}) - Q(c^j) - Q(c^i) + Q(c_a)$, where c^i, c^j and c^{ij} denote c_a with coordinate i, coordinate j, and both advanced.)

This is usually invoked as a caveat and then assumed away, but here the interactions are not negligible: the raster term with source support left free grows twenty-five-fold across the capacity ladder and collapses to approximately zero once support is held fixed (Section 5). The response is to change the estimand rather than the ordering — $\Delta_j \Delta_i Q$, the quantity that spoils the decomposition, is itself order-free, nuisance-invariant and measurable.

Two estimands. Fixing every coordinate except the training raster and measuring at a common test condition t gives

$$G(t) \;=\; Q\big(f_{4\mathrm{K}},\, t\big) - Q\big(f_{1\mathrm{K}},\, t\big), \qquad f_R = \mathcal{A}\big(\mathcal{D}_R,\, \mathcal{K},\, \mathcal{C},\, \mathcal{B}\big). \tag{3}$$

Crucially t is one tier for both arms; scoring each arm on its own tier, the default in practice, compares two different problems and is not a contrast at all. Varying t as well yields the second estimand, the *interaction span*

$$I \;=\; G(t_{4\mathrm{K}}) - G(t_{1\mathrm{K}}) \;=\; \Delta_R\, \Delta_t\, Q, \tag{4}$$

the same mixed second difference that Proposition 1 identified as the obstruction, now taken deliberately. The two estimands answer different questions: $G(t) > 0$ says only that the 4K arm is better on tier t, whereas I is large exactly when the *preference* between training rasters moves with the target resolution.

Collect the arm-by-tier scores into $\mathbf{Q} \in \mathbb{R}^{m\times n}$, $\mathbf{Q}_{Rt} = Q(f_R, t)$, and let $\mathcal{S}_{m\times n} = \{\mathbf{A} : \mathbf{A}_{Rt} = \alpha(t) + \beta(R)\}$ be the subspace of additively separable tables, in which the tier contributes an intrinsic difficulty and the arm a constant offset, with no interaction between them.

Proposition 2 (The interaction span is the unique separability-blind statistic)**.** $\dim \mathcal{S}_{m\times n} = m + n - 1$*, so the space of linear functionals annihilating* $\mathcal{S}_{m\times n}$ *has dimension* $(m-1)(n-1)$*. For* $m = n = 2$ *that space is one-dimensional and is spanned by the mixed second difference:* every *linear statistic of a* 2×2 *arm-by-tier table that is invariant to additively separable nuisance is a scalar multiple of* I*. Writing* $\mathbf{Q} = \mathbf{A} + \mathbf{\Psi}$ *with* $\mathbf{A} \in \mathcal{S}$ *gives* $I = \Delta_R \Delta_t \mathbf{\Psi}$*, which depends on the non-separable part alone. (Proof in Appendix B.)*

The dimension count is not decoration: with two rasters and two tiers there is exactly one interaction number to report, so I is not one summary among several we happened to prefer, and the full 3×3 transfer matrix of Section 6 carries $(3-1)(3-1) = 4$ independent interaction coordinates, of which I is the corner. Invariance is also what licenses comparing I across corpora whose operating points differ by several decibels: bicubic upsampling scores $35.98\,\mathrm{dB}$ on the RAISE 4K tier against $28.07\,\mathrm{dB}$ on DIV8K, so absolute Q, and even G at a single tier, are not comparable across sources, while the mixed difference removes exactly the terms responsible.

Scope of causal language. One limit is worth stating before the experiments rather than after. We identify a difference between data regimes; we do not claim to have characterised the boundary between them, and Section 7 reports six candidate predictors that fail to locate it. The second limit — that raster cannot be moved with source support, scene count and scene coverage all held, so no single configuration is neutral — is a property of the design space rather than of our design; Section 4 states it as equation 5 and describes how we bracket it.

4 Protocol

A design that identifies the raster term must hold every other coordinate of c fixed *by construction* rather than by assumption, since equation 2 shows that an unheld coordinate does not merely add a bias but makes the raster term order-dependent. We take the coordinates in turn; Figure 2 summarises the result.

The scene distribution: same-source pyramids. $\mathcal{X}$ is held by never letting two arms see different photographs: every tier is one resampling of the *same* original, no tier is derived from another, geometry is anchored on the low-resolution side as $|R| = s \cdot |R_{\mathrm{LR}}|$, and originals that do not reach the largest tier are discarded rather than upsampled. Scene content, camera, optics and capture are then identical across arms by construction, which is what turns G into an estimand.

The degradation operator, written down rather than imported. $\mathcal{K}$ is a single implementation shared across all tiers of all corpora — MATLAB-equivalent bicubic resampling with antialiasing, committed to the codebase rather than called from an image library. Section 5.1 is the reason: an operator that is assumed rather than declared is not a fixed coordinate, and getting it wrong costs $3.54\,\mathrm{dB}$.

The budget, and the crop that hides inside it. $\mathcal{B}$ and $\mathcal{C}$ are held by identical architecture, loss, batch size, crop size, schedule and step count within any comparison; only the corpus differs, and fixed steps make the exposure $E = T \cdot B \cdot hw$ equal by construction. One coordinate resists this. A fixed pixel crop is not a fixed crop policy: the same 192 px window is a fifth of a scene at 1K and a twentieth at 4K, so it contains two towers and a rose window on one tier and part of one arch on the other (Figure 9). Equalising crop size and equalising field of view are therefore different interventions and cannot both be satisfied. We equalise crop size, report the field of view alongside S, N and E, and measure the residual cost: $0.048\,\mathrm{dB}$ with source support free, $0.009\,\mathrm{dB}$ once it is matched (Table 4).

The coordinate that is left, and why it cannot be held alone. One coordinate stays entangled with R and no protocol can disentangle it. Writing a for the fraction of each scene's area an arm may draw from,

$$S \;=\; N\, a\, |R|, \tag{5}$$

binds four quantities at once, so varying $|R|$ *requires* releasing at least one of S, N and a: "the raster effect with everything else held fixed" is not an under-measured quantity, it is not a quantity. The constraint is scale-free, which is why it reappears inside the crop — a fixed pixel window covers s^2 times less scene at the higher tier, so matching crop size and matching field of view are the patch-level analogues of releasing a and releasing S, incompatible for the same reason.

Releasing each in turn gives exactly three designs and we run all of them (Table 1). *Unmatched* releases S, giving both arms every scene in full, which is what a conventional dataset comparison silently adopts; *support-matched* releases N, subsampling the higher-raster arm until the budgets agree; *iso-scene* releases a, keeping every scene but taking from each a window of the lower tier's pixel count. *Iso-pixel* in the results is the support-matched design swept across several budgets. A real raster advantage should survive whichever resource is released; one that appears under a single release identifies the resource responsible.

What counts as no effect, decided in advance. A null is only meaningful against a stated margin, so we fix one from measured variability rather than by choosing a round number. Three sources of noise enter a contrast and are never pooled — seed, matched subset, and iso-scene window placement — and sampling the source dominates seed noise by an order of magnitude (Appendix J). We take $\pm 0.05\,\mathrm{dB}$, about two and a half subset standard deviations and a shift of 1.2% in relative mean-squared error, as the interval within which an arm difference counts as no effect: a statement about this apparatus, not a universal threshold.

Keeping the measurement from manufacturing the effect. Train and test are separated at the scene level before any sampling, and for RAISE by capture group, since under a random split 36 of 117 test images shared a session with a training image. The RAW-to-RGB conversion is pinned in full, with explicitly *no* sharpening, denoising or auto-exposure. Every arm is scored on one common test tier and afterwards on all tiers, under PSNR, SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018a) on two backbones whose disagreements we report rather than resolve (Appendix J).

5 Apparent 4K Gains Are Explained by Source Support

5.1 The confound is consequential, not merely conceivable

One coordinate already decides a published comparison. A native-4K benchmark, 4KLSDB (Zhu et al., 2026), releases its low-resolution images alongside its high-resolution ones, so the pairing can be audited: the released pairs were produced by bicubic resampling *without* antialiasing followed by JPEG at quality 95, which recovers the distributed file *bit for bit* on all 100 pairs we obtained, where the antialiased convention reproduces only 11.4% of pixels exactly. The metadata labels the degradation `bicubic`. Isolating that factor in a 2×2 over training and evaluation operators costs $3.54\,\mathrm{dB}$, and the full input mismatch accounts for $5.98\,\mathrm{dB}$, larger than the $4.68\,\mathrm{dB}$ the benchmark reports after switching to native-4K data. We do not claim this explains that improvement, only that a quantity larger than the reported gain is attributable to a factor the comparison did not hold fixed (Appendix C).

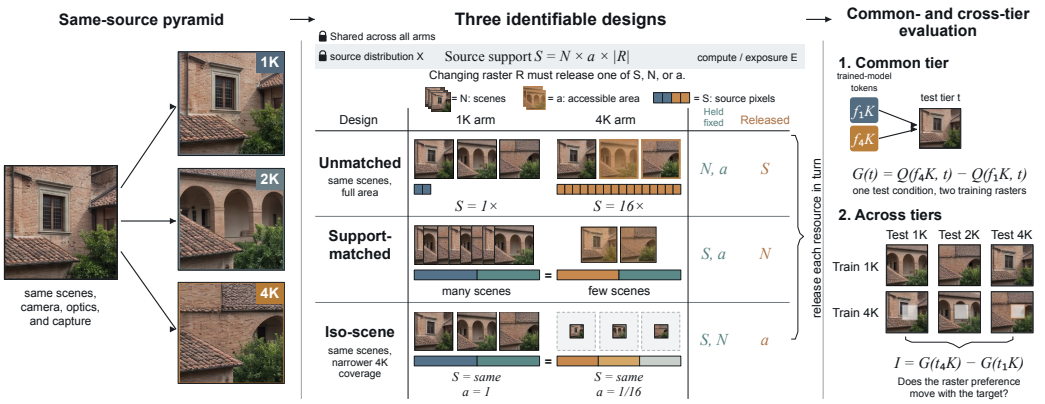

Figure 2: The protocol. **Same-source pyramid:** every tier is one resampling of the same original, so only the raster differs. **Three identifiable designs:** equation 5 binds four quantities, so moving the raster must release one of S, N or a; we run all three rather than calling one neutral. **Evaluation:** both arms on one common tier, which is what makes $G(t)$ a contrast, then across tiers to obtain I.

Table 1: The three resolutions of the budget constraint equation 5 on DIV8K, scored on the common test tier; positive favours the 4K-trained arm. Each row holds two of the three quantities raster is entangled with and releases the third. Only the row releasing source support shows a 4K advantage, and only it grows with capacity. Dashes are cells we did not run.

			EDSR, $\times 4$, at capacity			$\times 8$	SwinIR
Configuration	Held fixed	Released	1.5M	10.8M	43M	1.5M	1.5M
unmatched	N, area	S ($16\times$)	$+0.0328$	$+0.3961$	$+0.8349$	$+0.1530$	–
support matched	S, area	N ($558 \to 35$)	-0.0773	-0.0283	-0.0720	-0.0222	-0.0952
support *and* N matched	S, N	area ($1/16$)	-0.0377	-0.0264	-0.0062	–	–

5.2 Capacity amplifies the uncontrolled gap; matching support removes it

Held at a common test tier and a common degradation, the raster contrast depends entirely on whether source support is free or fixed. With both arms drawing on all 558 scenes, so that the 4K arm carries sixteen times the source pixels, the contrast grows steeply with capacity, from 0.033 dB at 1.5M parameters to 0.835 dB at 43M: read on its own, a compelling story about larger models exploiting a native-4K prior. Equalising the source-pixel budget — 321 against 322 megapixels, which costs the 4K arm all but 35 of its 558 scenes — removes it at every capacity, with four of the six controlled measurements inside the ± 0.05 dB equivalence interval and the two that are not (-0.077 and -0.072) missing it in the direction that disfavours the 4K arm (Table 1, Figure 3). The pattern survives a larger scale factor: at $\times 8$ the unmatched contrast is $+0.153$ dB and the matched contrast -0.022 dB (seed sd 0.0045), run at one capacity, so the claim is that the effect disappears at three capacities at $\times 4$ and at one at $\times 8$, not that a capacity-by-scale grid was swept.

The unmatched column at 43M needs a qualification: a fixed step budget is not a neutral control once capacity and data volume both vary, because it over-trains whichever arm has less data, and scored at each arm's own optimum that contrast falls from 0.835 dB to roughly 0.44 dB (Appendix D). We report both, the first being what the convention produces and the second what the comparison should have measured; their difference is another instance of Proposition 1. The matched columns are unaffected, both arms there holding equal data.

Perceptual metrics agree on the pattern but not the levels (Table 2): LPIPS-Alex prefers the 1K-trained arm in *every* one of the six cells, and the two backbones disagree in sign in four of them, so we make no perceptual claim in either direction. Matching S matches pixels and not information; an arm given the 4K pixel count without the 4K content loses 0.68 dB (Appendix E). Under a perceptual-adversarial objective, compared at matched PSNR, the null holds at the distortion end and a small but reproducible 4K advantage appears at the perceptual end (Appendix F).

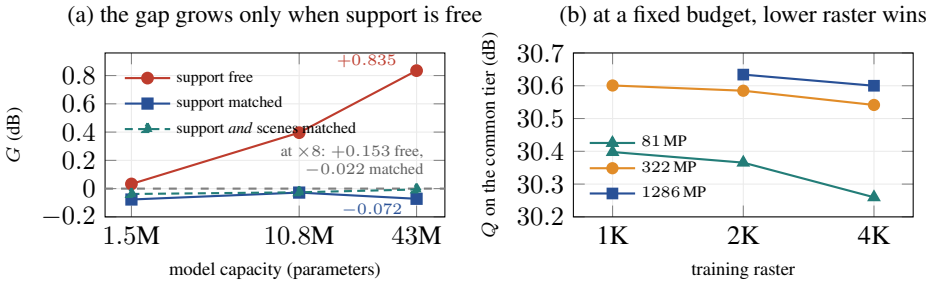

Figure 3: Two views of the same conclusion. (a) With source support free the raster contrast grows from 0.033 to 0.835 dB across the ladder, reading as a native-4K prior larger models exploit; matched, it sits below zero everywhere with no trend. (b) Within every pixel-budget contour, quality falls as the budget moves from many low-raster scenes to few high-raster ones. All arms on one common tier.

Table 2: Every arm of the source-support control under all four metrics, DIV8K at ×4 on the common test tier, mean of $n = 2$ seeds. G rows are signed so positive favours the 4K-trained arm in every column, LPIPS included. No metric shows a consistent advantage for either arm: the two LPIPS backbones disagree in sign in four of the six cells.

		source support free				**source support matched**			
Capacity	Train	PSNR ↑	SSIM ↑	LP-A ↓	LP-V ↓	PSNR ↑	SSIM ↑	LP-A ↓	LP-V ↓
1.5M	1K	30.5969	0.8318	0.2561	0.3110	30.5989	0.8317	0.2560	0.3110
(16 × 64)	4K	30.6296	0.8319	0.2620	0.3112	30.5217	0.8294	0.2619	0.3110
	G	**+0.0328**	+0.0002	−0.0059	−0.0001	**−0.0773**	−0.0022	−0.0058	+0.0001
10.8M	1K	30.5056	0.8331	0.2397	0.3062	30.5023	0.8331	0.2400	0.3064
(32 × 128)	4K	30.9017	0.8406	0.2460	0.3020	30.4740	0.8319	0.2415	0.3060
	G	**+0.3961**	+0.0075	−0.0063	+0.0042	**−0.0283**	−0.0012	−0.0015	+0.0003
43M	1K	30.1438	0.8262	0.2213	0.3021	30.1439	0.8262	0.2211	0.3020
(32 × 256)	4K	30.9787	0.8430	0.2412	0.2994	30.0720	0.8248	0.2239	0.3036
	G	**+0.8349**	+0.0168	−0.0199	+0.0027	**−0.0720**	−0.0014	−0.0028	−0.0016

5.3 Allocating a fixed pixel budget

If extra source pixels rather than raster carry the benefit, then at a fixed budget the raster should not matter. Holding the source-pixel budget approximately constant at three levels and varying the raster (Table 10 in Appendix K), higher raster never wins: the ordering is 1K > 2K > 4K in every contour, so a fixed acquisition or storage budget is better spent on more scenes at lower raster. The 2K–4K residual falls from 0.1057 dB to 0.0434 dB to 0.0340 dB as the budget grows, flattening rather than continuing geometrically, and the 1K–2K ordering sits inside the noise: what the data support is that 4K is lowest, not a ranking among the lower rasters.

5.4 Holding the scene count as well: the third resolution of the constraint

The support-matched configuration invites the sharpest objection: equalising S at fixed a forced the 4K arm from 558 scenes to 35, so a near-zero result is equally consistent with a raster effect cancelled by the loss of 523 scenes. By equation 5 no design settles this, so we resolve the constraint the remaining way. Releasing a keeps every scene and takes from each 4K scene a window of its own 1K dimensions — the same pixel count, $1/16$ of the area, at four times the density — so both arms hold 558 scenes and 321.3 MP, with patch pools within 0.2%. Windows are drawn per scene rather than centred, which would inherit photographic centre bias.

The cancellation reading predicts that restoring the scene count surfaces the raster effect. It does not (Table 1): at 1.5M, returning all 523 scenes is worth +0.0396 dB and leaves the contrast at −0.0377 dB. At 43M, where the uncontrolled effect is largest, the controlled contrast is smallest —

Table 3: Interaction span $I = G(t_{4K}) - G(t_{1K})$ and the per-tier contrasts it is built from, source-pixel matched at 1.5M, ordered by I. Daggers are our interventions on a source; D7000 and D90 are the two camera bodies RAISE was shot on, cut out of it separately. The last row is the top band's share of total magnitude in the 4K tier relative to the 1K tier, the statistic Appendix H shows does not predict the span.

	DIV8K	DIV8K† headroom	4KLSDB	FiveK (RAW)	RAISE 1920†	D90 body	RAISE 3840	D7000 body	RAISE web†
$G(t_{1K})$	-0.113	-0.132	-0.330	-0.364	-0.098	-0.371	-0.219	-0.226	-0.341
$G(t_{2K})$	-0.099	-0.098	-0.245	-0.225	0.071	-0.241	0.056	0.101	-0.196
$G(t_{4K})$	-0.077	-0.017	-0.096	0.097	0.373	0.144	0.556	0.616	0.562
I	0.036	0.114	0.234	0.461	0.471	0.515	0.775	0.842	0.904
Δ high-band (%)	-5	-41	$+4$	-17	-19	-31	-28	-32	–

$-0.0062\,\mathrm{dB}$ against a draw-to-draw spread of 0.0079 over three sets of window positions, smaller than the design's own sampling nuisance. Neither matched configuration shows a capacity trend. The two matched rows also hold *both* budget and raster, differing only in whether the budget buys 35 scenes whole or 558 at a sixteenth each, and the second wins at every capacity, putting the ordering of Section 5.3 on the coverage axis too.

The new arm sees $1/16$ of each scene, so the objection can be restated against it and no experiment closes the loop. What the pair establishes is that the two releases free *different* resources — scene count in one, area in the other — and neither uncovers a raster advantage at any capacity. For the cancellation account to survive, one raster benefit would have to be offset by scene loss in the first design and coverage loss in the second, at magnitudes that vary independently with capacity while the raster change stays identical.

6 How Far Resolution Dependence Varies

6.1 The interaction varies by source, and within one

Section 5 does not generalise. Repeating the matched-support design on three sources and two deliberately modified ones, scoring every arm on all three test tiers, yields interaction spans differing by more than an order of magnitude (Table 3, Figure 4). On DIV8K the training-raster preference is flat in the test tier; on RAW-derived RAISE it reverses sign, from $-0.219\,\mathrm{dB}$ at the 1K tier to $+0.556\,\mathrm{dB}$ at the 4K tier. FiveK, a separate project with different bodies, photographers and RAW pipeline, reverses as well, $I = 0.461$ with a seed spread of 0.008 at the crossing tier.

The eight spans progress smoothly with no gap, so we call this quantitative heterogeneity rather than two regimes; drawing the line at whether G changes sign would sort sources by their offset, the nuisance Proposition 2 removes. But we stop short of a continuum: a subset standard deviation of 0.020 separates four levels while leaving 0.461, 0.471 and 0.515 indistinguishable, precisely the neighbours such a claim would rest on, so mixture and threshold accounts remain open.

The variation is finer-grained than a dataset. Pyramids built separately from RAISE's two camera bodies give $I = 0.842$ and 0.515 — both reversing, but differing by more than the distance from DIV8K to 4KLSDB, within one collection processed one way by one group. A corpus-level span is therefore a mixture: RAISE's 0.775 is what a 72:28 blend of those bodies produces.

6.2 The interaction survives scene count, architecture, metric and split

Four perturbations that could each have removed the effect do not. Matching source support forces the 4K arm to 38 scenes against the 1K arm's 600, so raster and scene count are collinear and the interaction could reflect scene allocation; releasing the budget so both arms hold all 600, raising capacity to 10.8M, replacing EDSR with a shifted-window transformer, and splitting at random rather than by capture session give spans of 0.806, 0.807, 0.598 and 0.741 against the matched 0.775 (Table 9 in Appendix K). Four draws of the 38-scene subset give $G(t_{4K})$ between 0.501 and 0.607 (subset sd 0.051). All four metrics keep the span positive — $0.776\,\mathrm{dB}$ in PSNR, 0.0073

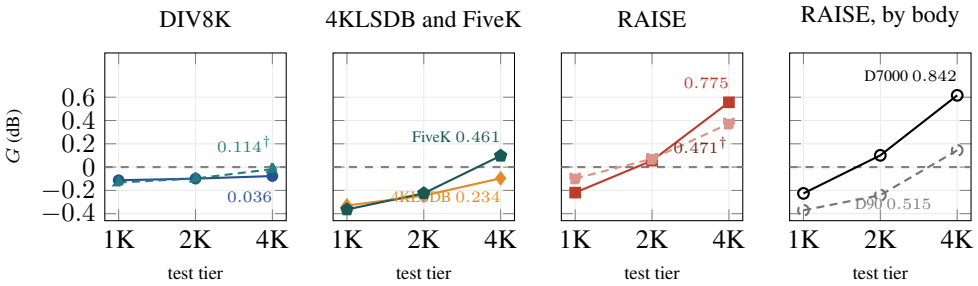

Figure 4: The same measurement across sources, dashed where we intervened. A flat line means the preferred raster does not depend on the target. Panels share a vertical axis, so by Proposition 2 the spreads are comparable despite operating points differing by several decibels. Numbers are I; the right panel splits RAISE by camera body, which both reverse.

in SSIM, 0.0302 in LPIPS-Alex, 0.0175 in LPIPS-VGG (Table 11) — though they disagree on the levels: at the 4K tier LPIPS-Alex still marginally prefers the 1K-trained arm, which is the distinction Proposition 2 draws, since a metric may carry an arbitrary offset in G and the span differences it away.

6.3 The reversal is distributional, and the preference tracks the target

A reversal could be produced by a handful of outliers pulling in opposite directions on the two tiers. It is not: on DIV8K with support matched, 91% of test images favour the 1K-trained arm, so the null is a small consistent preference rather than an average of opposing effects, and on RAISE the same pair of models is preferred by 79% of images at the 1K tier and 94% at the 4K tier — nearly every image changes its verdict (Figure 10 in Appendix K).

At the larger capacity the full transfer matrix is diagonally dominant, each training tier being the best choice for its own test tier (Appendix K). Strict dominance is a property of the 10.8M model, while the systematic shift of the raster preference across test tiers is present at both capacities, so we report the weaker statement — the preference moves with the target resolution — as the finding.

7 What Fails to Predict the Regime

Given a new corpus, can one tell in advance which regime it is in? We cannot: six candidate predictors fail, exploratory search included, so this bounds what we could find rather than what exists (Appendix H). Two are worth naming. A hand-crafted statistic fit four variants with a rank correlation of -1.000 and then came in at $+3.62$ on a held-out fifth, against an acceptance interval of $+10$ to $+17$ recorded before measuring. And spectral headroom, the most natural candidate, predicts nothing: the two RAISE bodies trace the same spectrum while their spans differ by 0.33 (Figure 7). Scalar summaries of marginal image statistics fail, as do local linear surrogates — consistent with the value of supervision concentrating in the first octave above the low-resolution Nyquist limit, precisely the structure such statistics discard.

8 Conclusion

The resolution written on a dataset is not a measure of the supervision it provides: a corpus swap also changes the source support, scene count, crop field of view and optimisation budget alongside the raster. Once the pixel budget is equalised the apparent advantage of native-4K data in a source-rich setting disappears across three capacities and two scale factors, does not return under any of the three ways the budget constraint can be resolved, and a fixed budget is better spent on more scenes than higher raster. On two unrelated RAW-derived corpora the same measurement reverses with the target resolution instead. What high-resolution supervision is worth is therefore an interaction between what a corpus supplies and the resolution at which a system is used: graded rather than binary, not predicted by any simple image statistic, but real, invariant to the nuisance that makes datasets incomparable, and cheap to measure.

AI USE STATEMENT

LLM-based tools were used under author supervision to assist with language polishing and manuscript drafting, as well as experimental implementation, execution, analysis, and visualization. The authors defined the research questions and experimental design, assessed the scientific claims, and revised the manuscript. Reported numerical results are traceable to the underlying experimental records. All authors take full responsibility for the final content and conclusions.

REPRODUCIBILITY STATEMENT

The degradation operator, the pyramid construction and the budget constraint that defines the three matched designs are specified in Section 4, with the construction rationale and measurement details in Appendix J. Every arm is determined by its corpus, tier, capacity, step budget and seed; the protocol holds architecture, loss, batch, crop, schedule and step count fixed within each comparison, and the values are given in the appendices. Seed and subset variability are reported separately rather than pooled, and the proof of Proposition 2 is in Appendix B. Code, per-arm data manifests and the result records behind every table are provided in the supplementary material, with path names anonymised, and will be released on publication.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference (BMVC)*, 2012.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 97–104, 2011.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. RAISE: A raw images dataset for digital image forensics. In *ACM Multimedia Systems Conference (MMSys)*, 2015.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

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

Netalee Efrat, Daniel Glasner, Alexander Apartsin, Boaz Nadler, and Anat Levin. Accurate blur models vs. image priors in single image super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *ACM Conference on Recommender Systems (RecSys)*, 2019.

Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. DIV8K: DIVerse 8K resolution image dataset. In *ICCV Workshops*, 2019b.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. In *International Conference on Learning Representations (ICLR)*, 2023.

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

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

Andreas Lugmayr, Martin Danelljan, and Radu Timofte. NTIRE 2020 challenge on real-world image super-resolution: Methods and results. In *CVPR Workshops*, 2020.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE International Conference on Computer Vision (ICCV)*, 2001.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020.

Go Ohtani, Ryu Tadokoro, Ryosuke Yamada, Yuki M. Asano, Iro Laina, Christian Rupprecht, Nakamasa Inoue, Rio Yokota, Hirokatsu Kataoka, and Yoshimitsu Aoki. Rethinking image super-resolution from training data perspectives. In *European Conference on Computer Vision (ECCV)*, 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV Workshops*, 2021.

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.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European Conference on Computer Vision (ECCV)*, 2020.

Jaejun Yoo, Namhyuk Ahn, and Kyung-Ah Sohn. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pp. 711–730, 2010.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, 2021a.

Kaihao Zhang, Dongxu Li, Wenhan Luo, Wenqi Ren, Björn Stenger, Wei Liu, Hongdong Li, and Ming-Hsuan Yang. Benchmarking ultra-high-definition image super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, 2021b.

Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, 2019.

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

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, 2018b.

Zihao Zhu, Kuan-Ru Huang, Zhaoming Xu, Renjie Li, Bo Wu, Ruizheng Bai, Mingyang Wu, Sayak Paul, and Zhengzhong Tu. 4KLSDB: A large-scale dataset for 4K image restoration and generation. *arXiv preprint arXiv:2605.24762*, 2026.

Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

A Discussion

Three things that "higher resolution" can mean. A corpus described as higher-resolution may be delivering any of three distinct goods, and the experiments separate them. It may supply more *source support*: more pixels to sample patches from, which is what the capacity ladder in Section 5 was actually measuring and which more scenes at a lower raster supply just as well. It may supply better *train–test alignment*: image statistics closer to those of the intended deployment resolution, which is what Section 6 isolates and which more scenes cannot substitute for. Or it may supply nothing beyond a larger number in the dataset description. These call for different decisions, and a single reported improvement does not distinguish them.

What to do with a fixed acquisition budget. The practical question a dataset builder faces is not whether 4K data is good but how to spend a bounded number of pixels. In the source-rich regime the answer from Table 10 is unambiguous: spend it on more scenes. Higher raster never won at any of the three budgets, and the cheaper option was at least as good everywhere. This inverts the usual intuition, on which acquiring at the highest available raster is a strictly dominant choice, and it matters because high-resolution acquisition is the expensive axis.

The RAISE result stops that from being a universal rule, and the exception is specific rather than vague. What changes is not the amount of data but whether the tiers remain statistically interchangeable; when they do not, the training raster that matches the deployment resolution is worth roughly half a decibel over one two octaves away, and no amount of extra low-raster scenes recovers it. The decision therefore depends on a relationship — between the source, the target resolution, and the resolution at which the system will be evaluated — rather than on a property of the data alone.

What a benchmark should report. Three of the numbers in this paper are large enough to reorder published comparisons on their own: $3.54\,\mathrm{dB}$ for a change of resampling operator, up to $0.835\,\mathrm{dB}$ for unmatched source support, and $0.048\,\mathrm{dB}$ for the crop protocol. None is usually reported. We suggest that a comparison across corpora of differing resolution state the degradation operator and its implementation; the source pixels and scene count of each arm, separately from the exposure; the split granularity, at scene or capture-session level; the acquisition device, since two camera bodies of one corpus differ in I by $0.33\,\mathrm{dB}$; the crop size and its field of view as a fraction of the source; the optimisation budget; and results on a common test tier rather than each arm on its own. Where the claim concerns resolution specifically, the cross-tier matrix costs only evaluation and settles in one table what a single-tier number cannot.

Why the source-rich side is empty. Appendix G measures the source-rich null rather than restating it. Because the marginal value of target bandwidth decays by about a factor of ten per octave, and because at a fixed scale factor both tiers present the model with the *same number of octaves above their own* Nyquist limit, raising the nominal raster supplies no additional relative band to learn from — which is what a matched G near zero should look like. This accounts for the source-rich regime only. It says nothing about why RAISE behaves differently, and we do not present the per-octave decay as a law.

Limitations. The controlled results come mainly from one architecture family, though both headline measurements replicate on a shifted-window transformer (Appendix I), and while capacity is varied over a factor of thirty we do not claim they transfer to generative or diffusion-based restoration, where the objective rewards different behaviour on exactly the unpredictable high-frequency content that Section 7 implicates. Real paired supervision is a structural gap rather than a missing experiment: a same-source pyramid requires one scene at several rasters, and a camera cannot photograph the same scene twice at different sensor resolutions, so our real-capture evidence comes from RAW sources demosaiced under a fixed pipeline rather than from natively paired captures. Perceptual evidence is inconclusive by construction — the two LPIPS backbones disagree in sign outside the matched cell, and we report that rather than selecting one. Most seeds are $n = 2$, so we quote effect sizes and spreads and avoid population statistics. And the ordering across sources is

only partly resolved: repeated subset draws put a standard deviation of about 0.02 on I, which separates four levels but leaves the three sources between 0.46 and 0.52 indistinguishable, so we report heterogeneity and not a continuum, and neither of our interventions crossed the point at which reversal appears. The upper end is populated by RAW-derived sources only: two sensors within RAISE and, independently, FiveK, all three reversing, against three web-scraped corpora that do not. Three reversing arms from two unrelated acquisitions is enough to rule out a single-sensor or single-project artefact, but not enough to say that being RAW-derived is what matters — that is a correlation across five sources, and Section 7 is the record of our failure to turn any such correlation into a predictor — including a direct test of this one, in which giving a RAW corpus the resize-and-JPEG history of a scraped one raised its span from 0.775 to 0.904 instead of collapsing it. FiveK is also our least precise source: its matched arm is 24 scenes against 31 and 35 elsewhere, and building it from 341 scenes rather than 490 moved its span from 0.461 to 0.595. The sign is stable, the magnitude is not, and we report the larger build. Section 7 shows that the obvious candidates for placing a corpus on that scale do not survive contact with a held-out source.

B Proof of Proposition 2

The map $(\alpha, \beta) \mapsto \alpha(t) + \beta(R)$ from $\mathbb{R}^n \oplus \mathbb{R}^m$ onto $\mathcal{S}_{m\times n}$ has kernel exactly the pairs $(c, -c)$ for constant c, so $\dim \mathcal{S}_{m\times n} = m + n - 1$ and its annihilator in the dual of $\mathbb{R}^{m\times n}$ has dimension $mn - (m + n - 1) = (m - 1)(n - 1)$. When $m = n = 2$ this is 1. Substituting a separable table into equation 4, each α appears once with each sign within a fixed tier and each β once with each sign within a fixed arm, so $\Delta_R\Delta_t$ annihilates $\mathcal{S}_{2\times 2}$ and, being non-zero, spans the annihilator. Linearity then gives the displayed identity,

$$I \;=\; \Delta_R\Delta_t\mathbf{\Psi} \;=\; \Psi_{4\mathrm{K},\,t_{4\mathrm{K}}} - \Psi_{1\mathrm{K},\,t_{4\mathrm{K}}} - \Psi_{4\mathrm{K},\,t_{1\mathrm{K}}} + \Psi_{1\mathrm{K},\,t_{1\mathrm{K}}}.$$

C Degradation Forensics

Before measuring anything, it is worth establishing that the confound in Section 3 is consequential in practice and not merely conceivable. We take one published native-4K benchmark as a case study, chosen because its low-resolution images are released alongside its high-resolution ones and the pairing can therefore be audited directly.

Reconstructing the released pairs shows that they were produced by bicubic resampling *without* antialiasing followed by JPEG compression at quality 95. This recovers the distributed low-resolution image *bit for bit* on all 100 released pairs we obtained — every pixel of every channel, not a close match — while the antialiased convention reproduces only 11.4% of pixels exactly on a high-frequency scene and differs by as much as 121 levels (Figure 5). The released metadata nonetheless labels the degradation `bicubic`. That operator differs from the antialiased resampling used throughout classical super-resolution, so a model trained under the classical convention meets an input distribution it has never seen. Isolating that single factor in a 2×2 over training and evaluation operators costs $3.54\,\mathrm{dB}$, and accounting for the full input mismatch, compression included, accounts for $5.98\,\mathrm{dB}$ — larger than the $4.68\,\mathrm{dB}$ improvement the benchmark reports after switching to native-4K data.

We are deliberately not claiming that this explains the reported improvement. Establishing that would require access to the original training runs, and the two numbers are measured under different conditions and cannot simply be subtracted. The claim is weaker and sufficient: a quantity larger than the reported gain is attributable to a factor that the comparison did not hold fixed, so the gain does not identify the value of native-4K supervision. That is the situation Equation equation 1 describes, and it motivates the controls that follow.

D Controls on the Matched Result

A null invites the objection that the experiment was insensitive rather than the effect absent. Table 4 collects the controls that address it; each row moves one aspect of the protocol and reports what happens to the matched contrast.

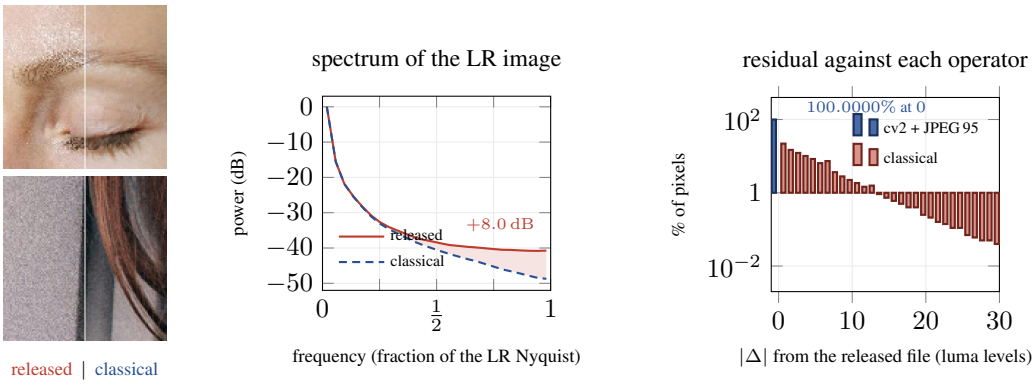

Figure 5: 4KLSDB's declared degradation is not the operator that made its data. **Left:** two crops split down the middle, released against the antialiased convention, at nearest-neighbour zoom. **Centre:** the released spectrum does not roll off — it flattens and sits 8 dB above the classical one at Nyquist, the signature of energy folded back by downsampling without a prefilter. **Right:** OpenCV `INTER_CUBIC` then JPEG 95 reproduces the released file bit for bit on all 100 pairs; the classical operator leaves 22% of pixels exact. Scenes are the three richest in energy above the low-resolution Nyquist limit.

Table 4: Controls on the matched-support result. Each row varies one aspect of the protocol; none changes the conclusion, and those that move G at all move it by less than the effect they were designed to rule out.

Control	What it varies	Effect
Optimisation length	40k vs. 200k steps, same arms	$-0.0757 \rightarrow -0.0727$
Crop protocol	fixed-pixel vs. FoV-matched crops	$-0.077 \rightarrow -0.068$
Capture realism	camera-like blur, noise and JPEG applied to both tiers	$+0.033 \rightarrow +0.005$
worst case	the same pipeline applied to the 1K tier only	$+0.078$
Resize provenance	$\times 4$ pyramid rebuilt from the 4K tier rather than the originals	$+0.0317$ vs. $+0.0328$
Perceptual metrics	SSIM, LPIPS-Alex, LPIPS-VGG on the matched cell	all ≤ 0
Implementation	Set5 / Set14 / BSD100 / Urban100	within 0.4–0.9 dB of published EDSR

Three of these deserve a sentence rather than a row. Extending training fivefold leaves the contrast intact, but *both* arms lose about 0.09 dB to over-fitting at 321 MP, so what is stable is the relative conclusion and not the absolute operating point. The crop protocol is worth 0.048 dB when source support is left free and only 0.009 dB once it is matched, which makes it a genuine reporting requirement for future comparisons rather than a robustness footnote; note also that FoV matching changes the low-resolution crop size and the target pixels per step along with the field of view, so it is not a one-sided stress test. And the two LPIPS backbones disagree in sign in four of the six cells of Table 2, so the claim we make is that no metric shows a consistent advantage for either arm, not that the perceptual evidence favours one.

On the four standard benchmarks (Bevilacqua et al., 2012; Zeyde et al., 2010; Martin et al., 2001; Huang et al., 2015), our models reach 32.09 dB on Set5 at 43M parameters against 32.46 dB for a fully trained EDSR, the gap being attributable to a 40k rather than 300k step budget and a different training corpus; the ordering across benchmarks and across capacities is correct. On these four test sets the 1K-trained and 4K-trained models are also within 0.06 dB of each other everywhere, including on Urban100, whose repetitive high-frequency structure is where a native-4K prior would be most expected to show.

E SOURCE SUPPORT COUNTS PIXELS, NOT INFORMATION

Equation equation 5 counts pixels. Every tier measured in the main text carries genuine detail at its own raster, so matching S has always matched information and pixels together, and a reader may fairly ask whether the matched null of Section 5 is an artefact of that bookkeeping rather than a fact about rasters.

A corpus of upscaled images separates the two. Resampling DIV8K's 1K tier up to the 4K tier's per-scene dimensions, with the same committed operator used everywhere else in this work, and deriving LR from the result, gives an arm that holds *exactly* the 4K tier's pixel count and the 1K tier's information. The construction matches the real 4K tier to 0.01% in source support (5136.6 MP against 5136.6 MP) and scene for scene in HR dimensions, while retaining between a twenty-sixth and a twenty-ninth of its energy above the 1K Nyquist limit.

Table 5: Arms holding a larger pixel count without the matching information, EDSR 16×64, 40k steps, three seeds each, scored on each corpus's own common test tier. "Area" is the pixel count relative to the 1K tier. The real-4K row on DIV8K reproduces the $+0.033$ of Table 1 to within the subset spread, which is the check that the comparison is wired correctly. The 4KLSDB 1K and real-4K arms are single-seed runs from the same 40k-step configuration.

Corpus	Training arm	area	S (MP)	Q (dB)	seed sd	G vs. 1K
DIV8K	1K	1	321.3	30.6012	0.0046	–
	1K upscaled 2×	4	1285.0	30.5819	0.0033	−0.0193
	1K upscaled 4×	16	5136.6	29.9229	0.0011	−0.6783
	real 4K	16	5136.6	30.6298	0.0033	+0.0286
4KLSDB	1K	1	312.9	34.1165	–	–
	1K upscaled 4×	16	5005.7	33.4446	0.0001	−0.6718
	real 4K	16	5005.7	34.0864	–	−0.0300

Capacity	1K	1K upscaled 4×	loss
1.5M	30.6012	29.9229	−0.6783
10.8M	30.4999	29.9125	−0.5874
43M	30.1458	29.9109	−0.2349
43M, at the 8k peak	30.6347	29.9496	−0.6852

If pixels alone drove the unmatched advantage, the upscaled arm would match the real 4K arm; if information drove it, the arm would match the 1K one. It does neither. Sixteen times the pixels with no added information is not worth zero, it is worth $-0.678\,\mathrm{dB}$ — a loss an order of magnitude larger than any residual Section 5 reports, and comparable in size to the uncontrolled advantage the paper sets out to explain.

We had predicted the upscaled arm would land *at* the 1K arm, the extra pixels contributing nothing. That prediction was wrong in a direction worth recording: upsampling does not merely fail to add information, it manufactures high-frequency content that the interpolator invented rather than the scene supplied, and an arm trained on it learns to reproduce that signature instead of real detail.

The cost is a threshold, not a slope. Doubling the linear size — four times the pixels — costs $0.019\,\mathrm{dB}$, which is inside the subset spread of 0.020 this paper uses as its equivalence margin. Quadrupling it costs $0.678\,\mathrm{dB}$. A constant rate of loss per unit of inflation would put the four-times arm at $-0.34\,\mathrm{dB}$; it sits $0.32\,\mathrm{dB}$ above that line. Mild upsampling is close to free and aggressive upsampling is not, which is what a model learning an invented signature rather than a scarcer one predicts: at $2\times$ most of the target spectrum is still content the scene supplied, while at $4\times$ three quarters of it is not.

The cost does not depend on the corpus. Repeating the $4\times$ construction on 4KLSDB gives $-0.672\,\mathrm{dB}$ against DIV8K's $-0.678\,\mathrm{dB}$. The two corpora differ by $3.5\,\mathrm{dB}$ in absolute operating point and by a factor of six in interaction span, and their upsampling losses agree to $0.007\,\mathrm{dB}$. Whatever else varies between sources — and Section 6 shows a great deal does — the price of inflating a corpus is a property of the operation rather than of the corpus.

Two things follow. Source support is a necessary bookkeeping device and not a sufficient one: S is blind to what the pixels contain, so a corpus can satisfy the matched-budget condition and still be worse than the tier it was built from. And where a corpus has been upscaled, the raster in its metadata is not merely uninformative but misleading — the right action is to resample down to the native tier before training rather than to spend the budget on the inflated one.

The loss is flat in capacity, once the budget is read correctly. Repeating the DIV8K contrast at 10.8M and 43M under the same 40k steps gives $-0.587\,\mathrm{dB}$ and $-0.235\,\mathrm{dB}$ against $-0.678\,\mathrm{dB}$ at 1.5M, which reads as the effect shrinking with scale. It is not. At 43M the fixed step budget over-trains the 1K arm — the effect Section 5 documents, and here it is sharper because the two arms differ sixteenfold in data: 73.4 scene-epochs against 4.6. Scored at the 8k peak instead, the 1K arm recovers to $30.6347\,\mathrm{dB}$ and the loss returns to $-0.685\,\mathrm{dB}$, so the capacity profile is -0.678, -0.587, -0.685 rather than a decline.

What the sweep does show is a ceiling. Across a thirtyfold increase in capacity the upscaled arm moves $0.012\,\mathrm{dB}$ in total, from $29.9229\,\mathrm{dB}$ to $29.9109\,\mathrm{dB}$ at 40k steps, and $29.9496\,\mathrm{dB}$ at the early stop. Thirty times the parameters buys an inflated corpus nothing, because what a model can extract is bounded by what the data contains: a practitioner who upscales a corpus and then scales the model to compensate is spending on the wrong axis.

Both corpora here are source-rich; we have not tested whether the loss appears on a corpus whose span reverses.

F The Matched Contrast Under a Perceptual-Adversarial Objective

Every result in the main text is for regression super-resolution under an L_1 loss. An objective that rewards plausible high-frequency content is the obvious place for native-4K supervision to help, so we repeated the DIV8K matched contrast with the ESRGAN recipe (Wang et al., 2018): L_1 at weight 10^{-2}, a VGG19 (Simonyan & Zisserman, 2015) `conv5_4` perceptual term, and a relativistic adversarial term (Jolicoeur-Martineau, 2019) at weight 5×10^{-3} against a spectrally normalised (Miyato et al., 2018) VGG-style critic. The generator, crop, batch and data pipeline are unchanged; the matched 4K arm is a fresh draw of 34 scenes holding 321.6 MP against the 1K tier's 321.3 MP, whose L_1 contrast of $-0.063\,\mathrm{dB}$ lies within the subset spread of the -0.077 reported in Table 1.

Why a single operating point is not a comparison. Trained with the same loss weights, the two arms land at different points on the perception–distortion trade-off (Blau & Michaeli, 2018): the adversarial term moved the 1K arm $3.14\,\mathrm{dB}$ off its L_1 model and the matched 4K arm only $1.79\,\mathrm{dB}$, because the critic's strength depends on the data it sees. Compared at that point the 4K arm is $1.29\,\mathrm{dB}$ higher in PSNR and worse in LPIPS-Alex — a difference of position on the curve, not of resolution. We therefore trace each arm's own curve and compare the arms at matched PSNR.

Protocol. For each arm and seed the adversarial generator is a 20k-step fine-tune of that arm's trained L_1 model, so the two share an exact common ancestor, and the curve is traced by network interpolation $\theta(\alpha) = (1-\alpha)\,\theta_{L_1} + \alpha\,\theta_{\mathrm{GAN}}$ at $\alpha \in \{0, 0.2, \ldots, 1\}$. No curve shows a barrier — no interpolant is worse than both endpoints on both PSNR and LPIPS — and PSNR falls monotonically along every curve. Perceptual metrics are compared on a common PSNR grid inside the overlap of all curves, $29.09\,\mathrm{dB}$ to $30.24\,\mathrm{dB}$.

Result. At the distortion end of the curve there is no 4K advantage, and in LPIPS-Alex the 1K arm is marginally ahead — the matched null of Section 5 carried over. At the perceptual end a small 4K advantage appears (Table 6). It is consistent in LPIPS-VGG and DISTS (Ding et al., 2022), where all twelve perceptual-end values favour the 4K arm, and weaker in LPIPS-Alex, where the three exceptions all come from one seed. Its largest value, 0.011 in LPIPS-VGG, is about a quarter of that metric's total movement along a curve.

Table 6: The DIV8K matched contrast under a perceptual-adversarial objective, compared at matched PSNR. Entries are the matched 4K arm minus the 1K arm, mean of two seeds; negative favours the 4K arm. The left column is the perceptual end of the overlap, the right column the distortion end. The last column counts, over two seeds and two code paths for the 1K arm, how many of the twelve values in the three perceptual-end columns favour the 4K arm.

	PSNR on the common grid (dB)					
Metric	29.09	29.38	29.67	29.95	30.24	favours 4K
LPIPS-Alex	-0.0052	-0.0084	-0.0008	$+0.0070$	$+0.0095$	9/12
LPIPS-VGG	-0.0081	-0.0114	-0.0077	-0.0031	-0.0001	12/12
DISTS	-0.0055	-0.0073	-0.0033	$+0.0006$	$+0.0020$	12/12

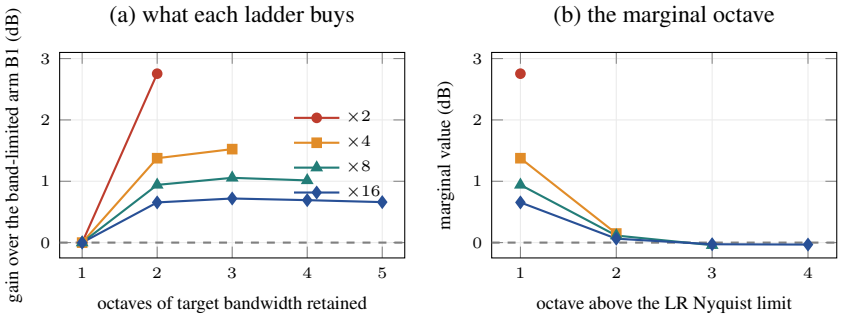

Figure 6: Band-limiting the training target, with the input and test set held fixed. Arm Bk retains k octaves above the low-resolution Nyquist limit, so B1 carries nothing the input does not already have and must reproduce bicubic — it does, to within $0.02\,\mathrm{dB}$, which is the construction check. (a) The ladder saturates almost immediately: whatever a scale factor is worth, the first octave delivers nearly all of it. (b) The same data as marginal values. Each octave is worth roughly a tenth of the one before, and beyond the second the contribution turns negative at $\times 8$ and $\times 16$.

The code path is not the cause. The 1K curve was built twice: once from the paper's L_1 models, trained with mixed precision, and once entirely in full precision on the same code as the 4K arm. Their L_1 endpoints agree to $0.002\,\mathrm{dB}$ in PSNR and 0.0006 in LPIPS-Alex. Across the fifteen seed-averaged cells of Table 6 the two curves differ by at most 0.0028; the only change of sign is LPIPS-VGG at the distortion-end column, where both values are within 0.003 of zero. None of the perceptual-end cells changes sign. The table reports the full-precision curve.

What this does and does not show. Matching source support removes the 4K advantage at distortion-oriented operating points under either objective. At perception-oriented ones, a small advantage survives the match, which suggests that resolution specialization, absent from the source-rich regime under L_1, may reappear in another form when the objective rewards generated detail. The evidence is two seeds on one corpus with an untuned adversarial recipe, the advantage is visible only in part of the overlap, and whether it grows further towards the perceptual end lies outside the PSNR range where the curves overlap.

G The Octave Profile

A null invites the suspicion that the measurement was insensitive, so it is worth showing that this one is the value the mechanism predicts. Restricting the training target to k octaves above the low-resolution Nyquist limit, with the input and the test set held fixed, isolates the marginal value of each octave of bandwidth (Figure 6). The first octave is worth roughly a decibel, the second about a tenth of that, and beyond the second the contribution turns negative — a pattern that holds at $\times 2$, $\times 4$, $\times 8$ and $\times 16$, with the ratio between successive octaves stable near 0.1.

This is what makes the matched result interpretable rather than merely small. At a fixed scale factor the two arms face the same number of octaves above their own Nyquist limits, so the higher raster

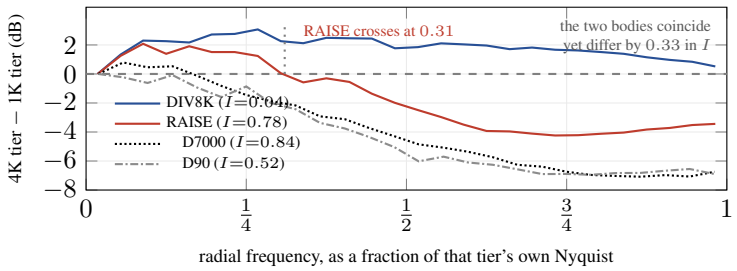

Figure 7: Each tier's spectrum relative to its own Nyquist limit, taken as the 4K tier minus the 1K tier so the comparison is of falloff shape and not detail level. DIV8K's tiers stay within 3 dB of each other; RAISE's cross at 0.31 and separate to −4.2 dB. The two RAISE camera bodies, cut out separately, trace the same curve while their spans differ by 0.33 — whatever the spectrum measures, it is not what sets the span.

supplies no additional *relative* bandwidth; almost all of the value available to either arm sits in the first octave, which both already have. A matched G near zero is therefore the predicted value, not a failure to resolve one.

H Predictors of the Regime That Fail

Section 7 summarises these six attempts; this appendix gives them in full.

Spectral headroom tracks an intervention and predicts nothing. The natural hypothesis is that specialization appears when the tiers stop being statistically interchangeable, and the simplest measure of that is how much high-band energy each tier retains relative to its own Nyquist limit. Within a source this works: raising RAISE's source headroom by lowering its top tier from 3840 to 1920 moves the high-band deficit from −28% to −19% and the span from 0.775 to 0.471, and reducing DIV8K's headroom in the opposite direction moves its span from 0.036 to 0.114. Both move the source and both point the right way. Neither is close to sufficient. Reducing DIV8K's high-band deficit to −41%, a stronger manipulation than RAISE's own −28%, recovers only about a tenth of the distance to RAISE's span, and the RAISE intervention recovers about two-fifths. Across sources the measure breaks outright: 4KLSDB's 4K tier is *sharper* than its 1K tier and its span is nonetheless six times DIV8K's (Table 3).

A hand-crafted statistic fits four points and fails a fifth. Because a model must predict the bicubic residual rather than the image, we next asked whether the tiers differ in how *predictable* that residual is, using the difference in its excess kurtosis between the top and bottom tiers as a proxy for sparse, structured detail against dense, noise-like detail. On the four variants then available this ordered the spans perfectly, with a rank correlation of −1.000 and R^2 above 0.99. Screening six candidate statistics on four points makes a perfect ordering roughly a coin flip, so before measuring the fifth we recorded a prediction: the held-out source should land at +13.3, with an acceptance interval of +10 to +17. It measured +3.62, and inverting the fit gives a predicted span of 0.537 against a measured 0.234. We report the interval and the failure rather than the fit.

Local surrogates: the structure is not linearly accessible. A statistic summarising single images may simply be the wrong object, since the span is a property of two *learning problems*. We therefore replaced it with a miniature of the experiment itself: ridge regression from a 7×7 low-resolution neighbourhood to the residual block it upsamples into, fitted on one tier and evaluated on the other, formed into the same matched-minus-mismatched contrast. On every variant the off-diagonal entries matched the diagonal to within 0.02 dB: the linear surrogate exhibits essentially no specialization at all. Whatever the tiers differ in, it is not recoverable by a local linear predictor.

Adding the cheapest useful non-linearity, a RAISR-style hash of local gradient orientation, strength and coherence with one filter per bucket, does produce specialization, and reproduces the extreme case closely (0.68 against a measured 0.775). It is not reliable: on RAISE with a 1920 top tier it returns −0.093 where the measurement is +0.471.

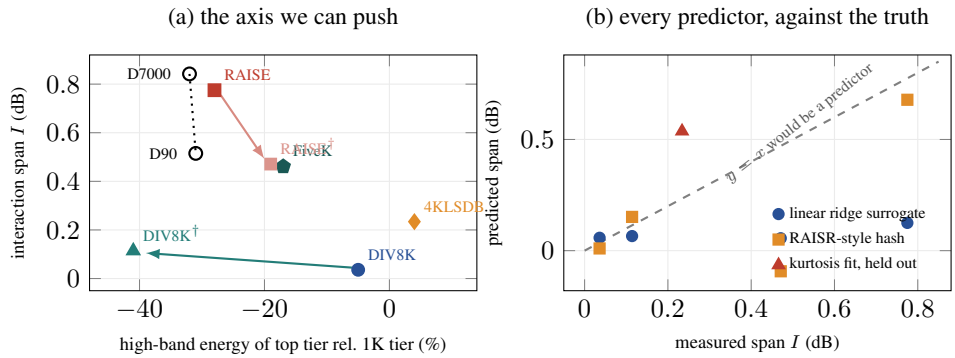

Figure 8: Why we report no regime predictor. (a) Interaction span against the spectral mismatch between a source's tiers. Daggers mark deliberate interventions, drawn as arrows from the unmodified source: reducing DIV8K's headroom moves it up the axis, raising RAISE's moves it down. Both act in the predicted direction, neither closes more than a fraction of the gap, and 4KLSDB sits well off any monotone relationship despite a *positive* mismatch. Open circles are the two RAISE camera bodies, joined by a dotted line: the same high-band deficit to within one point, spans 0.33 dB apart. (b) Predicted against measured span for the three quantitative candidates. The linear surrogate is flat — it predicts almost the same span whatever the truth — the hash tracks the extreme case and then returns a negative prediction for a source measured at $+0.471$, and the kurtosis fit, extrapolated to the held-out source, overshoots by more than a factor of two.

A fifth attempt, on the cleanest contrast available. Every failure above compares across corpora, which leaves the escape that the sources differ in too many ways at once. RAISE removes that escape: it was shot on two camera bodies, a Nikon D7000 (16.2 MP) and a D90 (12.3 MP), by the same photographers in the same period, and we can therefore build two disjoint same-source pyramids that differ in the sensor and in nothing else — identical demosaicing, identical degradation, identical grouped-split procedure, 500 training scenes and a 31-scene matched arm each.

Both reverse, which is the result Section 6 needs. But they reverse by very different amounts, and that is what defeats a fifth predictor. The two bodies sit at different distances from the 4K tier — the D7000 original is $1.28\times$ the 3840 target, the D90 only $1.12\times$ — so a source running out of resolution at its top tier should specialize more on the D90. The opposite happens: $I = 0.842$ on the D7000 against 0.515 on the D90, a gap larger than the entire distance from DIV8K to 4KLSDB, against a seed spread of 0.014. The spectral measure fares no better here than the sensor geometry does, and for a sharper reason: the two bodies have *the same* high-band deficit, -32% and -31%, so on that axis they are the same source and their spans should coincide. They differ by 0.33 dB (Figure 8a, the vertical pair).

Two things follow. Whatever sets the span varies at the granularity of a camera body rather than of a corpus, so "source" is not the natural unit for it — a corpus-level span such as RAISE's 0.775 is a mixture, here of two bodies in a 72:28 ratio. And a predictor that is blind to a difference this large, on the most controlled contrast we can construct, is not a predictor.

A sixth attempt, this one causal. The sources separate perfectly on one axis we have not yet tested: every corpus that reverses is demosaiced by us from camera RAW, and every corpus that does not is web-scraped, its originals already carrying an unknown resize and JPEG history. That is exactly the shape of evidence that misled us once already, so instead of asserting it we intervened on it. Taking RAISE, which reverses, we gave its originals the processing history a scraped corpus already has — the operator identified forensically in Section 5.1, OpenCV `INTER_CUBIC` at half scale and back, which antialiases in neither direction, followed by JPEG quality 95 — applied at the top tier and propagated, since a scraped original is damaged before any tier is derived from it. Split geometry, matched arms and test set are unchanged, and the damaged pyramid matches the clean one in total pixels to 0.2%.

If prior processing history is what suppresses specialization, the span should collapse towards DIV8K's 0.036. It does not collapse at all. It *rises*, from 0.775 to 0.904, moving away from the

Table 7: RAISE transfer matrices, source-pixel matched, scene-grouped split, PSNR in dB, mean of two seeds. Column maxima in bold. At 10.8M they lie on the diagonal; at 1.5M the 1K test column goes to the 2K arm instead.

	1.5M (16×64)			10.8M (32×128)		
Train \ Test	1K	2K	4K	1K	2K	4K
1K	33.228	36.105	40.021	**33.573**	36.505	40.346
2K	**33.295**	**36.413**	40.501	33.550	**36.679**	40.766
4K	33.009	36.161	**40.577**	33.250	36.450	**40.830**

web-scraped corpora rather than towards them, with G running -0.341, -0.196, $+0.562$ across the tiers.

This is the only attempt in which we changed one property of one source and read off the consequence, and it is correspondingly the most informative failure. The correlation between being RAW-derived and reversing is real across five sources and survives it; the obvious mechanism behind that correlation does not. Whatever makes a corpus resolution-specialized, it is not the absence of a resize-and-JPEG history, and a predictor built on acquisition provenance would have the sign backwards.

A caveat on the spectral measure. "High-band energy" names more than one statistic in this literature and in our own code, and the choices are not innocuous. The column used above is the top band's share of total *magnitude* on a 768 px centre window; computing the same share from *power* on a 512 px window reverses its sign on DIV8K, from -5% to $+22\%$, because the analysis window rather than the magnitude convention drives the difference. RAISE is negative and large under every variant we tried, so the cross-source failure reported here does not turn on the choice; the within-source ordering for DIV8K does. We report the measure as defined above, flag that it is not robust to the analysis window, and note that this fragility is itself a reason not to build a regime predictor on it.

I Replication on a Second Architecture

Everything in the main text uses one residual-CNN family, varied only in depth and width, which leaves open whether the results are properties of super-resolution or of EDSR. We therefore repeated the two headline measurements on SwinIR, a shifted-window transformer, at matched support and the same 40k-step budget.

Table 8: Both headline results under a change of architecture. PSNR on the common tier, matched support, one seed.

	SwinIR	EDSR
DIV8K, matched, common tier 4K (G)	-0.0952	-0.0773
RAISE $G(t_{1K})$	-0.2862	-0.2190
RAISE $G(t_{2K})$	-0.1003	$+0.0560$
RAISE $G(t_{4K})$	$+0.3113$	$+0.5560$
RAISE I	$+0.5975$	$+0.7750$

Both survive. On DIV8K the matched contrast stays small and negative, and the 4K arm gains nothing. On RAISE the preference still reverses with the target resolution, over a span of 0.598 against EDSR's 0.775.

One difference is worth pointing out because it illustrates Proposition 2 rather than contradicting anything. The tier at which G crosses zero moves: SwinIR is still negative at the 2K tier where EDSR has turned positive. The level of G carries architecture-dependent nuisance; the span, which the proposition identifies as the invariant part, is comparable across the two. This is one more reason not to classify a source by whether its curve crosses zero, as Section 6 declines to do.

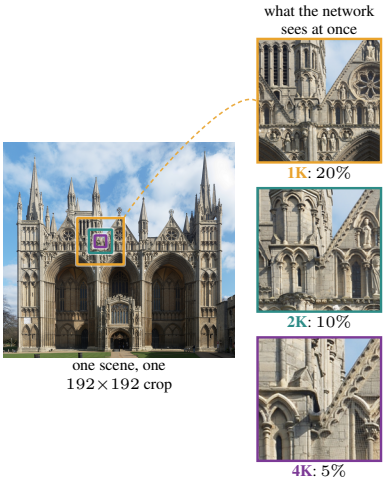

Figure 9: A fixed crop size is a different sampling instrument on each tier. Left: the region a 192 px crop covers, at true relative scale for each tier. Right: the crops themselves, each 192×192 *real* pixels — the 1K crop holds two towers and a rose window, the 4K crop part of one arch. Holding "crop size" constant across a raster change silently changes the problem.

J Protocol Rationale and Measurement Details

Why the pyramid is built the way it is. Anchoring geometry on the low-resolution side, $|R| = s \cdot |R_{\mathrm{LR}}|$, makes divisibility exact at every scale factor and removes rounding differences that would otherwise accumulate across tiers. Originals that do not reach the largest tier are discarded rather than upsampled, since upsampling one arm would reintroduce the very asymmetry the pyramid exists to remove. No tier is derived from another, so none inherits a second resize.

Why the operator is committed rather than imported. Calling a library resampler leaves the degradation hostage to a version bump that can silently redefine it. Section 5.1 is the concrete case: a published benchmark's declared operator turns out not to be the one that produced its data, and evaluating a model under the wrong one of the two costs $3.54\,\mathrm{dB}$ — larger than any effect this paper reports.

Why fixed steps, not fixed epochs. Fixing the number of optimisation steps makes the exposure $E = T \cdot B \cdot hw$ identical across arms by construction, so a difference between arms is attributable to what they *could* have sampled rather than to how much they did. Fixing epochs instead would confound the two.

Variance components. Retraining an arm moves G by a seed standard deviation of 0.0004 to 0.0136 across capacities; redrawing the matched subset at fixed size moves it by 0.020 on DIV8K and 0.017 on FiveK, over three independent draws each; and redrawing the window positions of the iso-scene arms moves it by 0.0079, over three draws. Sampling the source therefore dominates seed noise by an order of magnitude, which is why the two are never pooled.

Evaluation. The RAW-to-RGB conversion is pinned in full — demosaicing, camera white balance, gamma, bit depth, and explicitly *no* sharpening, denoising or auto-exposure — as each would reintroduce the class of artefact the protocol exists to exclude. We report PSNR on the luminance channel with a border shave equal to the scale factor, alongside SSIM and LPIPS under both AlexNet and VGG backbones. Seeds index training-level variability and subset draws index sampling variability; the two are never pooled, and with two seeds per cell we quote effect sizes and spreads rather than population statistics. Same-tier self-evaluation is never used as a contrast, since by Section 3 it compares two different problems.

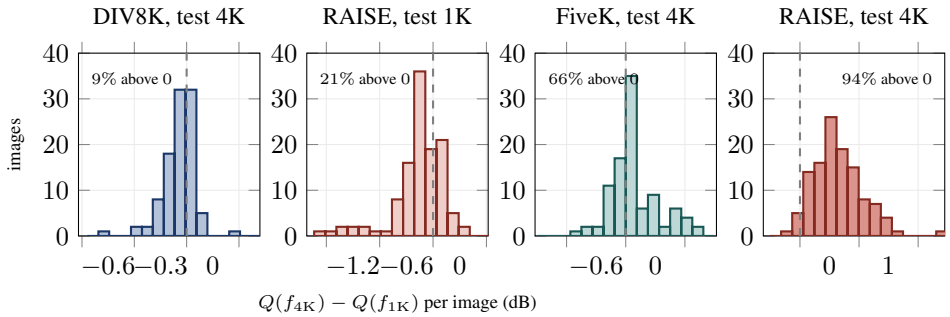

Figure 10: Per-image paired differences, not means. With support matched on DIV8K the preference is small but one-sided, so the near-zero G is not a cancellation. The centre panels show the *same* pair of RAISE models on two tiers, and FiveK at its 4K tier: the reversal shifts the entire distribution, 79% of images preferring the 1K-trained arm at RAISE's 1K tier against 94% the 4K-trained arm at its 4K tier.

Table 9: The interaction on RAISE under four perturbations that could each have removed it. The unmatched row does *not* have equal source support and therefore speaks only to scene allocation, not to Section 5; the last two rows change the architecture and the split rule. Dashes mark quantities a row does not define.

Setting	Scenes (1K / 4K)	$G(t_{1K})$	$G(t_{2K})$	$G(t_{4K})$	I
support matched	600/38	−0.219	0.056	0.556	0.775
support matched, 10.8M	600/38	−0.323	−0.055	0.484	0.807
support unmatched	**600/600**	−0.114	0.171	0.692	0.806
SwinIR, support matched	600/38	−0.286	−0.100	0.311	0.598
random rather than scene-grouped split	600/38	–	–	–	0.741

K Full Tables

The main text carries the contrast each of these tables was built to show; the versions here add the metrics, scene counts and spreads behind them.

Table 10: Iso-pixel contours on DIV8K, 1.5M parameters, $\times 4$. Within-cell spread is subset sd, except for the two full-set cells (321 MP at 1K and 1286 MP at 2K) where it is seed sd; the two are different quantities and are not pooled.

	~ 81 MP			~ 322 MP			~ 1286 MP		
Raster	N	Q (dB)	sd	N	Q (dB)	sd	N	Q (dB)	sd
1K	139	30.3978	0.0084	558	30.6008	0.0046	–	–	–
2K	35	30.3655	0.0200	140	30.5848	0.0038	558	30.6340	0.0036
4K	9	30.2598	0.0334	35	30.5414	0.0200	140	30.6000	0.0056

Table 11: RAISE at 1.5M under all four metrics, source-pixel matched, scene-grouped split, mean of two seeds. Best training raster per tier in bold. I is signed so that positive means the 4K-trained arm gains relative to the 1K-trained arm as the test tier rises; it is positive in every metric, although the metrics disagree on which arm leads at the 4K tier.

	test tier 1K			test tier 4K			
Metric	train 1K	train 2K	train 4K	train 1K	train 2K	train 4K	I
PSNR ↑	33.228	**33.295**	33.009	40.021	40.501	**40.577**	+0.7750
SSIM ↑	0.8628	**0.8656**	0.8600	0.9466	0.9506	**0.9511**	+0.0073
LPIPS-A ↓	**0.2253**	0.2374	0.2574	0.1627	**0.1566**	0.1646	+0.0302
LPIPS-V ↓	0.2768	**0.2742**	0.2828	0.2541	**0.2426**	0.2427	+0.0175