DOWNSTREAM TASK SCORES IN SUPER-RESOLUTION MEASURE THE PIPELINE, NOT THE PIXELS ALONE

Anonymous authors
Paper under double-blind review

ABSTRACT

Super-resolution (SR) is increasingly evaluated by downstream task scores, but these scores depend on the reconstructed images, the downstream model, and its decision rule. We audit how decision rules affect the performance and rankings attributed to SR. In remote-sensing segmentation, we hold reconstruction pixels, model weights, and continuous score maps fixed while comparing native decisions with a global threshold selected on validation data. Several SR pipelines score zero under the native rule even though their score maps retain discrimination above chance, but below that on matched high-resolution (HR) inputs. With HR-trained models, threshold selection nearly matches the best performance among the tested global thresholds, yet leaves a substantial gap to HR under matched conditions. Threshold gains persist when downstream models are fine-tuned on SR images. Large utility gains need not change method rankings, while changing the decision rule can alter which method ranks first. Controlled photography and microscopy analyses show that decision policies and target definitions can reverse comparisons, and audits of published downstream benchmarks show the same decision sensitivity, while the gap to matched references remains larger than the threshold gain from the default operating point. These findings motivate four reporting requirements: the decision budget, the matched reference, the target and cohort definition, and the claim scope. Together, they make clear what a downstream score measures and which comparisons it supports.

1 INTRODUCTION

Super-resolution (SR) is increasingly evaluated not only by pixel or perceptual fidelity, but also by whether reconstructed images support downstream vision tasks. In domains such as remote sensing and microscopy, reconstructions are fed directly to task models, with performance measured using metrics like F_1, IoU, or accuracy (Dai et al., 2016; Haris et al., 2021; Shermeyer & Van Etten, 2019; Rabbi et al., 2020; Li et al., 2026). This shift is valuable because it tests whether an image supports real-world tasks rather than whether it merely resembles a reference, complementing standard fidelity and perceptual metrics whose objectives can conflict (Blau & Michaeli, 2018). However, a downstream task score is produced by a compound pipeline comprising reconstruction, inference by a downstream model, which we call the *observer*, and a decision rule. While scores measured under a fixed observer and rule support comparisons under those specific deployment conditions, interpreting them as evidence of general reconstruction quality requires isolating what the observer and the decision rule contributed.

Even when the observer, labels, and task metric are fixed, a thresholded score reflects both the discrimination in continuous scores and the location of the decision threshold. A poor hard-decision score can therefore arise from degraded discrimination, an unsuitable threshold, or both. The score itself cannot distinguish between them. Operating points in image quality and calibration (Barrett et al., 2015; Provost & Fawcett, 2001; Guo et al., 2017; Kuppers et al., 2020; Wang et al., 2023), ranking metrics (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015), and frozen models under distribution shift (Recht et al., 2019; Taori et al., 2020) have each been studied independently. Yet downstream SR benchmarks sit at their intersection, attributing final task scores solely to upstream SR models even though reconstructions alter only inputs while scores reflect compound pipelines. We therefore ask how significantly decision policies alter absolute utility and comparative conclusions for fixed reconstructions, and what gap to matched high-resolution (HR) inputs remains after threshold selection.

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.

Our design fixes reconstruction pixels, observer weights, and continuous score maps while varying only the decision rule. Across five remote-sensing task–model settings from the GeoSR-Bench benchmark (Li et al., 2026), we evaluate six SR pipelines by comparing their native decision rules with a single global threshold selected on validation data. Thresholds are selected separately for each input arm, including HR, under the same selection policy. A shared observer trained on high-resolution (HR) imagery provides matched-HR comparisons within each setting; we then repeat the threshold intervention with released observers fine-tuned on each pipeline's super-resolved outputs. Re-evaluations of two released downstream benchmarks extend the decision audit to object detection and camouflaged object detection. Finally, controlled analyses on camera photography with City100 (Chen et al., 2019) and fluorescence microscopy with BioSR (Qiao et al., 2021) examine sensitivity to decision policies and target definitions under handcrafted feature responses.

Two empirical findings motivate this audit. First, with an HR-trained SegFormer on USBuildings, four SR pipelines yield a native F_1 score of exactly zero despite retaining substantial continuous discrimination, where mean within-image AUROC remains above 0.80 while matched HR achieves 0.99. A hard-decision score of zero therefore conflates absent discrimination with preserved, but mis-thresholded, discrimination. Second, even after downstream observers are fine-tuned on super-resolved imagery, selecting a validation threshold yields mean gains of 0.100–0.169 F_1, with positive point estimates across all 30 pipeline–observer pairs. Decision sensitivity thus survives downstream adaptation and extends beyond near-zero scores. Furthermore, absolute utility and method ordering decouple. While full-split rankings for UNet on RoadDetection remain perfectly stable, SegFormer on RoadDetection alters the top-ranked method in point estimates without statistically resolving the new leading pair.

Our contribution is a way to read a downstream score by its parts, providing a controlled audit of decision sensitivity and a reporting framework for downstream evaluation. Combining decision interventions with continuous-score diagnostics and matched references separates four distinct phenomena that a single score conflates, namely threshold failure, degraded discrimination, threshold headroom, and the remaining gap to matched HR. The separation carries beyond remote sensing. On fixed camera and microscopy images, the decision policy and the target definition can each reverse a comparison, and on two published downstream benchmarks the decision effect replicates. We distill these findings into four reporting requirements that govern decision budget, matched reference, target and cohort definition, and claim scope. Together, they let claims about super-resolution rest on what reconstructions achieve under a stated observer and evaluation protocol, rather than on how arbitrary thresholds were chosen.

2 Reporting Requirements for Downstream Evaluation

A frozen downstream model, the observer, maps a reconstruction to continuous scores, a decision rule converts those scores into hard predictions, and the task metric compares those predictions with labels. A task score alone therefore cannot identify which stage limits performance. We codify what a downstream evaluation must report into an *evaluation contract* comprising four reporting requirements that govern decision budget, matched reference, target and cohort definition, and claim scope. These requirements make decision effects identifiable and reconstruction comparisons interpretable. Table 1 lists the information each requirement asks a report to state and the misreading it prevents.

R1: Decision budget. Substituting super-resolution outputs into an unchanged downstream system evaluates compatibility at zero additional adaptation budget, whereas selecting a threshold on validation data assesses achievable task utility when validation labels may guide the decision rule. Fine-tuning the observer on super-resolved images adds model adaptation to this budget. Selecting a threshold on test labels instead gives an in-sample optimum within the candidate family, not a deployable selection procedure. A reported score should state its decision rule, supervision and model updates, candidate family, and selection split.

R2: Matched reference. To attribute a performance gap between high-resolution and super-resolved inputs to the reconstruction rather than to a change in the evaluation pipeline, the high-resolution reference must use the same observer, input dimensions and resampling, threshold-selection policy, and evaluation cohort. Matching threshold policies means matching the selection procedure and supervision budget rather than enforcing identical numerical cutoffs across divergent score distributions. Without such a reference, paired gains within a pipeline remain valid but do not establish recovery of high-resolution performance. Collapsed hard decisions such as zero F_1 should be assessed alongside continuous-score diagnostics on both super-resolved and high-resolution inputs. A comparison with high-resolution imagery should report these reference conditions explicitly.

Table 1: The evaluation contract. For each requirement, what a report should state and the misreading it prevents, with the section that tests it.

Requirement	What to report	Misreading it prevents
R1: Decision budget	The decision rule, the budget spent on it, the candidate family, and the selection split	Crediting the reconstruction with utility that the threshold supplied, or reading a score tuned on test labels as deployable, Section 4.2
R2: Matched reference	An HR reference under the same observer, input dimensions and resampling, threshold policy, and cohort	Reading a gap to an HR reference run under a different observer, input size, threshold policy, or cohort as a gap in image quality, Section 4.2
R3: Target & cohort definition	Target construction, evaluated cohort, and aggregation unit	Attributing to the method a change that comes from the target definition, the cohort, or the aggregation unit, Section 4.5
R4: Claim scope	For utility, the conditions it was measured under; for rankings, paired differences with intervals over independent units	Reading a utility gain as a ranking change, or a top-1 change whose paired interval spans zero as a reversal, Section 4.3

R3: Target and cohort definition. Because decision rules operate against specific target formulations, modifying ground-truth definitions can invert the empirical ranking of competing decision policies. Cohort filters that exclude test tiles overlapping validation boundaries change foreground prevalence and shift the trivial baseline, allowing relative position against high-resolution ceilings to diverge substantially even while raw F_1 scores remain nearly unchanged. Pooling counts across the dataset and averaging within images also weight scenes differently, even when all images have the same dimensions, because pooled metrics weight each scene through its confusion counts whereas per-image averaging weights each scene equally. A reported score should state how the target was constructed, which cohort was evaluated, and the unit of aggregation.

R4: Claim scope. Utility gains, changes in empirical ordering, and statistically resolved superiority are distinct findings. A large utility gain can leave method rankings intact, while the leading method can change without resolving the new leading pair. Claims of task utility must identify the governing budget, observer, and cohort, whereas method comparisons should report paired differences between complete pipelines and their sensitivity to the decision rule. Inference beyond the evaluated sample requires uncertainty evaluated at a sampling unit that accounts for the relevant spatial or grouping dependence.

HR-relative position. R2 places utility on a common baseline-to-HR scale across settings. Let $F_{\mathrm{all}+} = 2\pi/(1+\pi)$ be the pooled F_1 of predicting every pixel positive at foreground prevalence π, and let F_{HR} be the pooled F_1 of the same frozen observer on HR input under the same dimensions, threshold policy, and cohort. We define $p = (F - F_{\mathrm{all}+})/(F_{\mathrm{HR}} - F_{\mathrm{all}+})$, assuming that $F_{\mathrm{HR}} > F_{\mathrm{all}+}$, a condition that holds strictly in every setting examined. Thus $p = 0$ at the all-positive baseline and $p = 1$ at the HR reference. This relative scale resolves misleading conclusions drawn from raw F_1 metrics. For example, SegFormer reaches a higher validation-selected F_1 on RoadDetection than on USBuildings, registering 0.386 against 0.354, yet occupies a lower relative position of 16.4% against 30.6% because the all-positive baseline on RoadDetection is nearly twice as high, as shown in Table 2. Within any single setting, shared anchors make p an increasing affine function of F_1, which preserves rankings. Across distinct cohorts, both anchors must be recomputed as required by R3. The position is intentionally left unclipped across its domain, allowing negative values to explicitly signal performance falling below the all-positive baseline, as occurs under native decision thresholds, while values exceeding one would indicate performance surpassing the empirical HR reference.

Scope of the audit. The primary audit fixes reconstruction pixels and observer weights in remote-sensing segmentation and varies only the decision rule. A second audit repeats the intervention on released pipeline-adapted observers, and a third on two published downstream benchmarks with their released observers. Controlled analyses vary the decision policy on the City100 benchmark (Chen et al., 2019), target construction on the BioSR dataset (Qiao et al., 2021), and cohort filters and score transformations on remote-sensing data. Together, these interventions test decision sensitivity and the evaluation choices that govern its interpretation. The outcome of interest is hard-decision utility.

3 Audit Design

We fix reconstruction pixels, observer weights, and score maps while comparing native decisions, validation-selected thresholds, and a test-label diagnostic optimum. The first two represent deployment budgets; the third measures remaining finite-grid headroom. Appendix A gives implementation details.

3.1 Decision Rules

An input arm is an SR pipeline, bicubic interpolation, or HR. The observer produces foreground and background logits ℓ_1, ℓ_0 with margin $z = \ell_1 - \ell_0 = \log(\hat{p}_1/\hat{p}_0)$ for a two-class softmax head. All rules use the same maps, and thresholds are selected separately for each arm and task–observer setting. The intervention therefore changes how an existing score distribution is converted into predictions, with neither the image nor the observer retrained between decisions.

The *native decision*, $\hat{y}_{\text{native}} = \mathbb{I}(z > 0)$, is two-class argmax with ties assigned to background, equivalent to cutting $\hat{p}_1 = \sigma(z)$ at 0.5. This plug-in MAP rule targets symmetric zero-one loss, not F_1: under calibrated probabilities the F_1-optimal threshold is $F_1^*/2 \leq 0.5$ (Lipton et al., 2014). We therefore measure native compatibility separately from validation-selected task utility.

The *validation-selected decision* fits one scalar on 160 labeled validation tiles:

$$c^* = \min \arg\max_{c \in \mathcal{C}^{\text{val}}} F_1^{\text{val}}(c), \qquad \hat{y}_{\text{val}} = \mathbb{I}(z > c^*). \tag{1}$$

Here $\mathcal{C}^{\text{val}}$ contains pooled validation-margin quantiles at 2001 uniformly spaced levels from 0 to 1. Ties favor the lower threshold and foreground recall. Weights remain fixed, no test labels enter selection, and c^* transfers unchanged to test data. We report $F_{\text{val}} = F_1^{\text{test}}(c^*)$ and $F_{\text{native}} = F_1^{\text{test}}(0)$. The search uses quantile levels rather than uniformly spaced logit increments, so it follows the empirical score distribution. The one-parameter budget changes only the operating point; it does not estimate a probability-calibration map or modify the observer.

The *diagnostic grid ceiling*, $F^*_{\text{grid}} = \max_{c \in \mathcal{G}} F_1^{\text{test}}(c)$, uses $\mathcal{G} = \mathcal{C}^{\text{val}} \cup \mathcal{C}^{\text{test}}$, with test-margin quantiles constructed analogously. Test labels identify this optimum, so it is a diagnostic, not a deployable selection rule. Since $c^* \in \mathcal{G}$, headroom $F^*_{\text{grid}} - F_{\text{val}} \geq 0$. Small headroom relative to the HR gap limits the contribution of suboptimal selection within this grid.

3.2 Tasks, Inputs, and Observers

USBuildings and RoadDetection (Li et al., 2026) each provide 160 validation and 400 test tiles at 1000×1000. We use released checkpoints trained for 16-fold super-resolution of Sentinel-2 imagery using NAIP reference images. The six pipelines are ATD (Zhang et al., 2024), BiDiffSR, CAMixerSR (Wang et al., 2024), CFAT (Ray et al., 2024), RGT (Chen et al., 2024), and SRNO (Wei & Zhang, 2023); BiDiffSR is supplied by the benchmark. Floating-point SR outputs are fixed. Bicubic interpolation operates in sensor radiometric space rather than the learned output space, making it a distinct input baseline; its limiting all-positive case is examined in Section 4.2. Released HR-trained UNet (Ronneberger et al., 2015), SegFormer (Xie et al., 2021), and Swin (Liu et al., 2021) evaluate USBuildings; UNet and SegFormer evaluate RoadDetection. SR outputs are 1024×1024, so matched HR is bilinearly resampled to that size. Logit maps are bilinearly returned to the 1000×1000 label grid before thresholding; binary masks are never interpolated. Thus the reference and SR images reach the observer at matching dimensions, while all hard predictions are evaluated on the same native ground-truth grid. Section 4.4 uses released pipeline-adapted observers on the same inputs and splits.

3.3 Metrics and Statistical Inference

Pooled foreground $F_1 = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$ sums counts across tiles. Pooled $\text{IoU} = F_1/(2 - F_1)$ is strictly increasing, so both metrics select identical thresholds and pipeline orderings. The derivative $2/(2 - F_1)^2$ is positive throughout $[0, 1]$. This equivalence does not extend to mean per-image IoU (Appendix G.1): the transform is strictly convex, and applying it before averaging need not preserve the ordering of pooled metrics. The aggregation unit therefore remains part of the evaluation protocol, even when the underlying pixel predictions are identical. Continuous discrimination is measured by tie-corrected, within-image AUROC, averaged equally across tiles; every tile contains both classes. All-positive F_1 supplies the prevalence-dependent lower anchor defined in Section 2.

Model-selection stability. We report tie-adjusted Kendall's τ_b and inversions among the six SR pipelines. A pair enters the inversion denominator only if both rules order it strictly; a tie under either rule counts as neither agreement nor inversion.

Uncertainty and spatial scope. We use 95% percentile intervals from $B = 2000$ paired bootstrap draws. RoadDetection resamples whole source groups, preserving dependence among their constituent tiles. USBuildings resamples individual tiles because only two regional grid cells are available; ignoring inter-tile spatial dependence may understate uncertainty. The two cells contain spatially neighboring tiles, so many tile draws do not constitute independent regional replications. We therefore examine rankings separately within each cell and test transfer on source groups absent from validation, rather than treating tile-level intervals as evidence of broad geographic generalization. Absolute utility and headroom intervals resample test units with the selected threshold fixed. Ranking and paired-difference intervals resample validation and test units, refitting the threshold on the fixed candidate grid in each draw. Arms share each draw, while groups present in both splits are resampled independently within each split. The first design conditions on the deployed decision policy; the second also represents variation in threshold selection. These intervals answer different questions and are reported accordingly (Appendix A).

4 Empirical Audit

Table 2 summarizes five settings. We test R2 through continuous diagnostics, R1–R2 through decision gains and HR gaps, R4 through ranking stability, and R1 under observer adaptation. Section 4.1 contrasts native scores with continuous discrimination; Section 4.2 separates recovered utility from the remaining HR gap; Section 4.3 measures changes in method ordering; and Section 4.4 repeats the decision intervention after observer fine-tuning. Section 4.5 extends these comparisons to other domains and tests target, cohort, and score-transformation sensitivity under R3.

Table 2: Threshold selection lifts every setting but stays far below matched HR. Test F_1 under the native cut (F_{native}) and the validation threshold (F_{val}), reported as means across the six SR pipelines. HR-relative position (p) is reported for the mean across pipelines, the range across the six individual arms, and bicubic interpolation. Precision spans the range of test precision across pipelines under the validation threshold, and τ_b measures rank agreement between native and validation orderings. All entries are point estimates.

		SR F_1		HR-relative position (p)					
Task	Model	F_{native}	F_{val}	mean	six arms	bicubic	HR F_{val}	Precision	τ_b
USBuildings	SegFormer	0.009	0.354	30.6%	28.5–34.5%	11.4%	0.821	0.254–0.302	−0.086
USBuildings	Swin	0.003	0.289	20.7%	17.0–25.1%	12.6%	0.829	0.191–0.252	0.467
USBuildings	UNet	0.005	0.195	6.9%	0.7–15.6%	0.0%	0.823	0.085–0.207	0.733
RoadDetection	SegFormer	0.042	0.386	16.4%	10.4–25.4%	0.0%	0.857	0.246–0.338	0.733
RoadDetection	UNet	0.003	0.467	30.7%	20.4–39.7%	8.5%	0.858	0.293–0.423	1.000

4.1 Native Decision Performance and Continuous Discrimination

In all five evaluation settings, the mean native F_1 across the six pipelines is at most 0.042, while the same frozen observers under the same native cut score at least 0.816 on HR inputs; Appendix B lists all five settings. The continuous scores nonetheless separate foreground from background: mean AUROC within images ranges from 0.609 for UNet to 0.813 for SegFormer on USBuildings, and individual learning-based SR arms span 0.563 to 0.825 in Table S2. Matched HR reaches 0.973 to 0.987 in the same settings, so discrimination is reduced but not lost.

SegFormer on USBuildings illustrates this most clearly. Under the default cut $z > 0$, four pipelines predict no positive pixels at all, namely ATD, CAMixerSR, RGT, and SRNO. Their pooled test F_1 is therefore exactly zero, yet on the same score maps their mean AUROC within images stays well above chance, from 0.804 to 0.824, Figure 1a. Matched HR reaches 0.986 under the same observer. Average precision within images, Appendix C, shows a larger deficit relative to matched HR than AUROC does.

For RGT in this setting, every pixel margin across all 400 test tiles falls below the default cut, the maximum being -0.095, so the native rule predicts no positive pixels. Yet the two class distributions remain partially separated, and the validation threshold $c^* = -6.734$ lands directly within their overlap, Figure 1b. The native threshold therefore fails to expose the discrimination present in these maps. Reading the hard decision together with continuous diagnostics separates this operating-point failure from the remaining gap to matched HR, as required by R2.

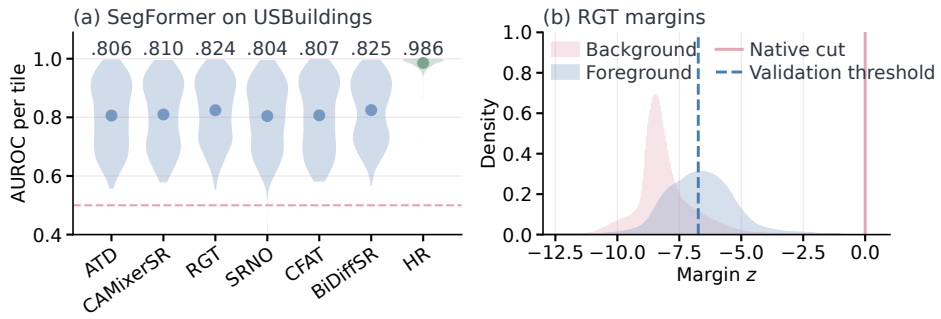

Figure 1: Zero native F_1 does not imply absent discrimination. (a) Within-image AUROC for SegFormer on USBuildings across all 400 test tiles, evaluated across six super-resolution pipelines alongside matched high-resolution imagery relative to the 0.5 chance baseline, where numerical annotations indicate tile means. (b) Margin densities of RGT foreground and background pixels, each normalized by class size, illustrating how the native cutoff $z = 0$ classifies all pixels as background whereas the validation-selected threshold c^* captures the overlapping score distributions.

4.2 Threshold Adaptation and the Matched-HR Gap

The surviving discrimination raises a practical question: how much task utility can a validation threshold recover from the same score maps? Across all five settings, six-pipeline mean F_1 rises to 0.195–0.467, including an increase from 0.003 to 0.467 for UNet on RoadDetection (Table 2; Figure S1, Appendix B). Remaining grid headroom is at most 0.0016 for a setting mean and 0.0075 for an individual arm (Appendix C). Four alternative rules, three applied per image, improve no setting mean; only $8/120$ rule–arm configurations improve, by at most 0.0071 (Appendix D). Thus the selected threshold recovers nearly all utility available on the evaluated grid, and the tested alternatives add little on average.

A substantial gap to HR nevertheless survives. Matched references score 0.821–0.858, well above the recovered SR performance. The reference itself depends on protocol: matching Swin's HR inputs to the 1024×1024 SR grid shifts its USBuildings score from 0.762 to 0.829. R2 therefore requires matched dimensions as well as matched observers. Lowering the threshold also admits false positives; for UNet on USBuildings, test precision is only 0.085–0.207.

The all-positive baseline clarifies what these gains mean under sparse foreground. All 30 SR arms exceed it in point estimates, but RGT under UNet on USBuildings clears it by only 0.0047, and intervals for RGT and CAMixerSR extend below it. Bicubic with SegFormer on RoadDetection reaches the limiting all-positive prediction, $p = 0$. Learned SR arms occupy just 0.7%–39.7% of the baseline-to-HR span (Appendix C).

The gains also transfer beyond the tuning sources: without refitting, thresholds retain at least 95.8% of the native-to-grid-best gain on 48 held-out-source RoadDetection tiles across all 12 pipeline–observer combinations (Appendix F). Taken together, these results locate two distinct sources of low task scores. The large native-to-validation gain comes from choosing an operating point on an unchanged score map. The much smaller remaining grid headroom shows that this choice has nearly exhausted the evaluated threshold budget. The persistent HR gap then measures the deficit that remains under matched evaluation conditions. Reading these quantities together distinguishes a successful decision adjustment from recovery of HR-level task performance.

4.3 Decision Sensitivity of Task Utility and Model Rankings

Each arm receives its own threshold, so all scores can improve without changing their order. Whether utility gains alter method selection is therefore an empirical question. UNet on RoadDetection provides the clearest separation: mean F_1 rises from 0.003 to 0.467 while ordering stays unchanged: $\tau_b = 1.000$, $0/15$ inversions, and RGT remains first under both rules in all 2000 bootstrap draws (Figure 2b). Utility gains need not change rankings.

SegFormer on the same task shows a different response: $2/15$ pairs invert ($\tau_b = 0.733$), and the full-split leader changes from BiDiffSR to RGT; the rules select different winners in 94.4% of resamples. BiDiffSR's native advantage is $+0.1311$ with interval $[+0.0975, +0.1489]$. After selection, RGT scores 0.4365 versus BiDiffSR's 0.4235, but the BiDiffSR-minus-RGT difference -0.0130 has interval $[-0.0382, +0.0029]$ (Figure 2a). The native lead disappears, while the two methods remain

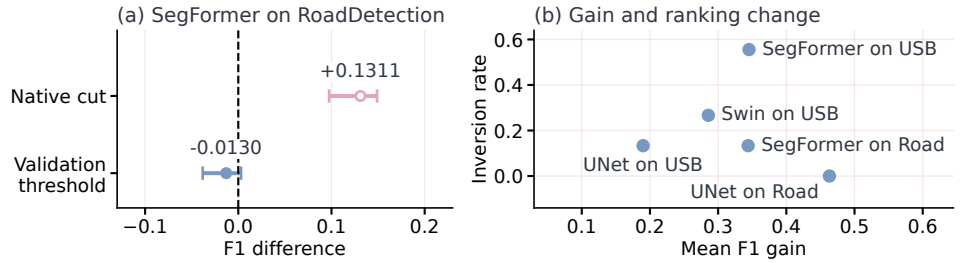

Figure 2: Downstream task utility gains decouple from model ranking stability. (a) Paired difference in F_1 between BiDiffSR and RGT for SegFormer on RoadDetection under the native cut and validation threshold, accompanied by 95% source-group bootstrap intervals from 2000 resamples, where the native interval excludes zero whereas the validation interval spans zero. (b) Mean F_1 gain across the six super-resolution pipelines plotted against the fraction of strictly orderable pairs that invert, where setting labels abbreviate USBuildings as USB and RoadDetection as Road.

statistically unresolved after threshold selection. The frequent change in the point-estimate winner therefore captures sensitivity of method selection to the decision rule.

This contrast extends across observers and spatial populations. Across settings, τ_b spans -0.086–1.000 and inversions $0/15$–$5/9$ (Appendix E). For SegFormer on USBuildings, four zero-scoring arms create six tied pairs, leaving nine orderable pairs. SegFormer and Swin change leaders in $2000/2000$ and $1997/2000$ tile resamples, respectively. Yet when thresholds are refitted within MGRS cells, both changes persist in the 318-tile cell and disappear in the 82-tile cell. UNet has no full-split leader change. The tile-bootstrap winner is therefore not necessarily the winner in both geographic regions. These results connect R4 to two separate checks: uncertainty in paired performance and sensitivity of the comparison to the evaluated population.

4.4 Threshold Sensitivity under Pipeline-Adapted Observers

The primary audit uses observers trained on HR, so the decision gains could depend on substituting SR into an unadapted system. We test this explanation using released observers fine-tuned on each pipeline's SR training images (Li et al., 2026). This gives 30 pairs over five settings; bicubic is excluded because no adapted observer was released. Reconstructions, splits, and each observer's weights stay fixed between decision rules (Appendix H).

Adaptation raises mean native F_1 from 0.003–0.042 to 0.363–0.507, yet threshold selection adds 0.100–0.169 in every setting, with positive gains in all 30 pairs (Figure 3). Excluding BiDiffSR, the largest-gain arm in every setting, still leaves mean gains of at least 0.063 (Table S10). The effect therefore persists across pipelines after the larger adaptation budget has been spent.

Rank correlations span 0.467–0.867; RGT yields first place to CAMixerSR in $2/5$ settings (USBuildings–SegFormer and Swin), with margins of 0.0042 and 0.0006. These small margins describe the empirical ordering; paired uncertainty was not estimated for this audit.

Observer adaptation and threshold selection are complementary budgets: adaptation adds 0.358–0.466 native F_1, followed by further threshold gains. Adapted SR means reach 0.532–0.607, versus 0.821–0.858 for HR inputs under HR-trained observers, differences of 0.251–0.292. This is an end-to-end system comparison across observer budgets, not a same-observer SR–HR contrast. Within each adapted observer, however, the positive threshold gain is measured on unchanged reconstructions and weights. Observer fine-tuning improves default compatibility without eliminating the separate contribution of the decision rule.

4.5 Cross-Domain Analyses and Evaluation Sensitivity

The remaining analyses ask which conclusions depend on the wider evaluation protocol. City100 photographs and BioSR microscopy use handcrafted responses, extending the audit beyond neural segmentation. We then vary score scales and cohorts in remote sensing and repeat the decision intervention on two published deep-vision benchmarks.

Decision policy and SR method comparisons. City100 (Chen et al., 2019) uses Shi–Tomasi corners (Shi & Tomasi, 1994) on fixed reconstructions (Appendix G). SmallSR *zero* and *standard* share architecture and losses, but *zero* masks high-frequency adversarial gradients during training. Their

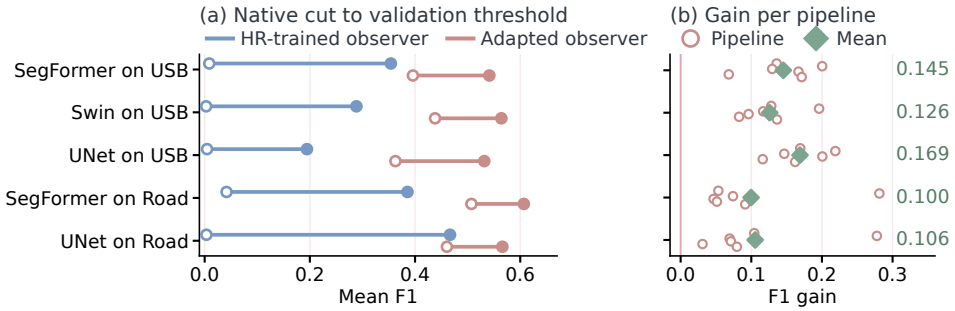

Figure 3: Threshold gains persist after observer adaptation across all five settings. (a) Mean F_1 across the six super-resolution pipelines under HR-trained and adapted observers, where each segment connects native and validation-selected performance for one observer condition. (b) Threshold gains for all 30 adapted pipeline-observer pairs, with setting-specific means annotated. Both panels report full-test-split point estimates.

paired F_1 contrast is -0.150 under a reference-anchored policy using test-HR response scales, versus $+0.013$ under a calibration-fitted global threshold; both intervals exclude zero. Retaining the reference policy but applying a calibration-fitted scale correction gives $+0.010$, also resolved (Figure S4a). The policy change reverses the supported method preference. The reference policy is diagnostic because it accesses test HR.

Target definition and decision-rule comparisons. On BioSR (Qiao et al., 2021), predictions and rules stay fixed. A parameter-free per-image 0.94-quantile rule beats a calibration-fitted global threshold by about 0.06 F_1 across three SmallSR losses when targets use each Sato-filtered HR image's 0.94 quantile (Sato et al., 1998). Under Otsu (Otsu, 1979) or hysteresis (Canny, 1986) targets, the same rule loses by 0.06–0.09, with intervals excluding zero; an absolute target leaves the difference unresolved (Figure S4b). The same predictions thus support opposite preferences among decision rules when the target changes. R3 makes that target dependence explicit, separating the quality of a reconstruction from compatibility between a decision policy and the labeling convention.

Score transformations and aggregation units. Global monotonic margin transformations across three USBuildings observers change pooled F_1 by up to 0.380 with a fixed numerical cut; refitting restores it within 10^{-6}. Image-specific scaling instead changes pooled rank area by up to 0.168 while preserving every per-image rank area (Table S7). Score-scale and aggregation invariance are therefore distinct.

Evaluation population and reference-relative standing. Removing spatial overlap with validation excludes $123/400$ USBuildings and $7/400$ RoadDetection tiles, without refitting thresholds (Table S6). USBuildings–SegFormer F_1 changes only from 0.3540 to 0.3543, but p rises from 30.6% to 34.4% under the retained cohort's lower prevalence and recomputed anchors. For UNet, F_1 drops 0.0259 while p rises from 6.9% to 8.3%. Cohort composition changes relative standing with pixels, weights, and thresholds fixed.

Published downstream benchmarks and reference dominance. EDTR (Kim et al., 2025) uses a fine-tuned Faster R-CNN on 5823 VOC2012 images. With official observers and code, two-fold cross-fitting at IoU 0.5 shifts cuts from 0.50 to 0.63–0.77, adding 0.0204, 0.0229, and 0.0170 micro-F_1 for HR, four-step EDTR, and SR4IR. This captures 87%–99% of exhaustive-grid headroom (Table S8). The benchmark reports continuous mAP above a 0.05 score floor; released operating cuts span 0.22–0.35 micro-F_1 within checkpoints. EDTR's 0.0229 gain from 0.50 is smaller than its remaining 0.0918 matched-HR gap, but its 0.3530 gain from 0.05 exceeds that gap. The release score floor and the 0.50 operating point answer different questions: gains from one must not be substituted for gains from the other. In particular, continuous mAP does not specify the hard-decision deployment rule audited here.

SR-COD (Wen et al., 2025) evaluates seven observers on 2012 COD10K images with fourfold bicubic downsampling. The adaptive cut $\min(2 \cdot \text{mean}, 1)$ trails fixed 0.50 in $43/49$ observer–input configurations, including clean inputs and both SwinIR training variants. Its mean IoU loss of 0.0123 exceeds the mean grid-optimal gain of 0.0009 by over thirteen times (Table S9). The leading SR method stays first across cuts; its gain is 0.0008 against a 0.0343 clean-reference deficit, over

forty times larger (Appendix G.1). These audits preserve the distinction between decision gains and remaining reference gaps.

5 Related Work

Downstream evaluation of super-resolution. Detection and segmentation benchmarks (Dai et al., 2016; Haris et al., 2021; Shermeyer & Van Etten, 2019; Rabbi et al., 2020; Li et al., 2026) complement distortion and perceptual measures (Blau & Michaeli, 2018; Zhang et al., 2018). The perception–distortion trade-off makes these task evaluations complementary to image similarity, but a downstream score introduces additional dependencies on the observer and decision rule. We quantify those dependencies on fixed reconstructions and interpret recovered utility against matched HR, rather than treating the terminal metric as a direct measure of pixels.

Foundations in decision theory. Task-based image quality depends on observers and operating points (Barrett et al., 2015; Provost & Fawcett, 2001); F_1-optimal thresholds depend on prevalence and score distributions (Lipton et al., 2014). Calibration (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017; Minderer et al., 2021; Kuppers et al., 2020) and ranking metrics (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015) characterize complementary score properties. Downstream SR places these familiar components in a cascade: an upstream reconstruction changes the input to a frozen observer whose scores feed a decision rule. Our audit combines continuous diagnostics, decision interventions, and HR references to distinguish threshold failure, discrimination loss, recoverable utility, and remaining deficits. Together, these established tools provide a controlled analysis of score attribution in SR.

Benchmark design and evaluation sensitivity. Metrics, aggregation, and benchmark construction can change conclusions (Maier-Hein et al., 2024; Reinke et al., 2024; Dehghani et al., 2021; Bouthillier et al., 2021; Musgrave et al., 2020; Lorenz et al., 2024; Kahl et al., 2024). In weakly supervised localization, controlling thresholds and validation supervision revised gains over CAM (Choe et al., 2020). We extend this discipline to downstream SR, where the reconstruction, observer, and rule jointly determine the final score. Our reporting requirements make decision budgets, matched references, targets and cohorts, and claim scope explicit, linking each choice to a measured change in how the result can be interpreted.

For SR evaluation, a central question is whether an advantage between reconstruction methods persists under equal decision budgets. A shared numerical threshold fixes the decision rule, but different reconstructions can shift the observer's score distributions differently. Allowing each method the same validation budget provides a complementary comparison. Reporting both shows which advantages depend on native compatibility and which persist after threshold selection. Our audit makes this distinction observable on fixed reconstructions, connecting benchmark rankings to the deployment conditions and adaptation resources under which those rankings are useful.

6 Conclusion

Downstream task scores measure reconstruction, observer, and decision rule jointly. Zero native scores can conceal discrimination; validation threshold selection recovers utility even after observer adaptation. In the HR-trained audit, little evaluated grid headroom remains, yet matched-HR gaps persist. Utility gains and method rankings need not move together, and leader changes need not establish superiority. Camera and microscopy experiments show resolved reversals from decision-policy or target changes. Published-benchmark audits further show that gain magnitude depends on the baseline: reference gaps dominate gains from 0.50, but not EDTR gains from the 0.05 floor. The practical implication is to match the reported score to the intended use: native performance evaluates direct substitution, while validation-selected performance evaluates a label-assisted deployment budget. Reporting both alongside matched references, explicit targets and cohorts, and paired ranking uncertainty makes these dependencies visible. A downstream score then supports a reconstruction comparison under stated conditions, with decision gains distinguished from recovery of HR-level task performance.

Reproducibility Statement

Section 3 and the appendices document the evaluation and resampling protocols. For the primary and pipeline-adapted remote-sensing audits, the supplementary material contains 16 CSV files covering per-arm results and confusion counts, HR references, decision-rule comparisons, ranking and resampling results, and tile-level split assignments. The README maps reported results to the relevant files and columns. Protocols and summary results for City100, BioSR, EDTR, and SR-COD are provided in Appendix G.

AI Use Statement

Generative AI tools were used to assist with language polishing, manuscript organization, and drafting scripts for data analysis and plotting. All experimental designs, scientific analyses, numerical results, and interpretations were conceived by the authors, who take full responsibility for the integrity, correctness, and final text of this work.

References

Harrison H. Barrett, Kyle J. Myers, Christoph Hoeschen, Matthew A. Kupinski, and Mark P. Little. Task-based measures of image quality and their relation to radiation dose and patient risk. *Physics in Medicine and Biology*, 60(2):R1–R75, 2015. https://doi.org/10.1088/0031-9155/60/2/R1.

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

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021.

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986. https://doi.org/10.1109/tpami.1986.4767851.

Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://doi.org/10.1109/cvpr.2019.00175.

Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. In *International Conference on Learning Representations (ICLR)*, 2024. https://doi.org/10.48550/arxiv.2303.06373.

Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. https://doi.org/10.1109/cvpr42600.2020.00320.

Dengxin Dai, Yujian Wang, Yuhua Chen, and Luc Van Gool. Is image super-resolution helpful for other vision tasks? In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 233–240, 2006.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In *International Conference on Neural Information Processing*, pp. 387–395. Springer, 2021.

Kim-Celine Kahl, Carsten T. Lüth, Maximilian Zenk, Klaus Maier-Hein, and Paul F. Jaeger. ValUES: A framework for systematic validation of uncertainty estimation in semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2024. https://doi.org/10.48550/arxiv.2401.08501.

Jaeha Kim, Junghun Oh, and Kyoung Mu Lee. Exploiting diffusion prior for task-driven image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. https://doi.org/10.48550/arxiv.2507.22459.

Fabian Kuppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 326–327, 2020.

Zhili Li, Kangyang Chai, Zhihao Wang, Xiaowei Jia, Yanhua Li, Gengchen Mai, Sergii Skakun, Dinesh Manocha, and Yiqun Xie. Beyond visual fidelity: Benchmarking super-resolution models for large-scale remote sensing imagery via downstream task integration. *arXiv preprint arXiv:2605.00310*, 2026.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Optimal thresholding of classifiers to maximize F_1 measure. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 8725 of *Lecture Notes in Computer Science*, pp. 225–239, 2014. https://doi.org/10.1007/978-3-662-44851-9_15.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Julian Lorenz, Alexander Pest, Daniel Kienzle, Katja Ludwig, and Rainer Lienhart. A fair ranking and new model for panoptic scene graph generation. In *European Conference on Computer Vision (ECCV)*, 2024. https://doi.org/10.1007/978-3-031-73030-6_9.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020. https://doi.org/10.1007/978-3-030-58595-2_41.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 625–632, 2005.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 1979. https://doi.org/10.1109/tsmc.1979.4310076.

Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 2001. https://doi.org/10.1023/a:1007601015854.

Chang Qiao, Di Li, Yuting Guo, Chong Liu, Tao Jiang, Qionghai Dai, and Dong Li. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature Methods*, 18, 2021. https://doi.org/10.1038/s41592-020-01048-5.

Jakaria Rabbi, Nilanjan Ray, Matthias Schubert, Subir Chowdhury, and Dennis Chao. Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network. *Remote Sensing*, 12(9):1432, 2020.

Abhisek Ray, Gaurav Kumar, and Maheshkumar H. Kolekar. CFAT: Unleashing triangular windows for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. https://doi.org/10.1109/cvpr52733.2024.02468.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019.

Annika Reinke, Minu D. Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Emre Kavur, Tim Rädsch, Carole H. Sudre, Laura Acion, Michela Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature Methods*, 21(2):182–194, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015.

Yoshinobu Sato, Shin Nakajima, Nobuyuki Shiraga, Hideki Atsumi, Shigeyuki Yoshida, Thomas Koller, Guido Gerig, and Ron Kikinis. Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images. *Medical Image Analysis*, 2(2):143–168, 1998. https://doi.org/10.1016/S1361-8415(98)80009-1.

Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *CVPR Workshops, EarthVision*, 2019.

Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994. https://doi.org/10.1109/CVPR.1994.323794.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yan Wang, Yi Liu, Shijie Zhao, Junlin Li, and Li Zhang. CAMixerSR: Only details need more "attention". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. https://doi.org/10.1109/cvpr52733.2024.02441.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Min Wei and Xuesong Zhang. Super-resolution neural operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. https://doi.org/10.1109/cvpr52729.2023.01750.

Juan Wen, Shupeng Cheng, Weiyan Hou, Luc Van Gool, and Radu Timofte. When super-resolution meets camouflaged object detection: A comparison study. *Computer Vision and Image Understanding*, 253:104321, 2025. https://doi.org/10.1016/j.cviu.2025.104321.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. https://doi.org/10.1109/cvpr52733.2024.00276.

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

A Additional Evaluation Details

Benchmark pipeline. The primary satellite audit uses GeoSR-Bench (Li et al., 2026), which pairs Sentinel-2 inputs with NAIP reference images for 16-fold super-resolution. Each Sentinel-2 tile contains 12 spectral bands of 60×60 raw digital counts, which are divided by 3500 and resized to 64×64. The super-resolution network outputs a four-band 1024×1024 reconstruction clamped to $[0, 1]$, which is passed directly to the frozen observer. High-resolution tiles are scaled by $1/255$ and bilinearly resampled from 1000×1000 to the matching 1024×1024 input resolution. All predicted logits are interpolated back onto the original 1000×1000 ground truth label grid. Pooled F_1 is calculated from aggregate pixel counts on this grid, whereas AUROC and AP are computed per tile and then averaged. A single HR-trained observer is evaluated across all eight arms in each primary setting; Appendix H examines adapted observers tailored to each pipeline.

Threshold grids and tie-breaking. For each arm and split, candidate thresholds are 2001 empirical quantiles of pooled pixel margins, equally spaced from probability zero to one via linear interpolation; duplicate quantile values are retained. The validation rule selects the smallest threshold maximizing pooled validation F_1. The diagnostic grid ceiling searches for the threshold that maximizes test F_1 over the union of candidates derived from both validation and test splits. Under the strict decision rule $z > c$, setting a threshold below the minimum test score produces an all-positive prediction, as observed for bicubic interpolation with SegFormer on RoadDetection. Table S2 reports the diagnostic headroom for each individual arm.

Continuous score diagnostics and rank statistics. We compute AUROC and average precision independently for each tile and average them across tiles. When computing pipeline rankings, arms with identical pooled F_1 scores are treated as tied rather than strictly ordered. This occurs under native decisions when multiple arms predict zero positive pixels. Tied pairs are excluded from both the numerator and denominator of the inversion fraction. If C and D denote concordant and discordant pairs, and T_0 and T_1 denote pairs tied only under native or only under validation decisions, then $\tau_b = (C - D)/\sqrt{(C + D + T_0)(C + D + T_1)}$, while the inversion fraction is $D/(C + D)$. Because of ties, these two metrics are not simple linear rescalings of each other.

Resampling. The primary audit uses two bootstrap designs, each drawing sampling units with replacement over $B = 2000$ replicates. For RoadDetection, the sampling units are source groups; drawing a group includes all of its constituent tiles, preserving natural group sizes. For USBuildings, sampling units are individual tiles. Sample sizes differ across splits; RoadDetection contains 52 validation and 67 test source groups, whereas USBuildings contains 160 validation and 400 test tiles. Groups present in both splits are resampled independently within each split. Each bootstrap draw is applied simultaneously across all arms in a setting to maintain strictly paired comparisons.

Confidence intervals for pipeline rankings, top-rank selections, and pairwise differences are computed by resampling both validation and test units, refitting c^* on the resampled validation data in each draw while keeping the candidate threshold grid fixed. In contrast, confidence intervals for absolute F_1 levels and diagnostic headroom resample only test units while holding c^* fixed at its full split value; these intervals are therefore conditional on the selected threshold. The held-out source analysis on the 48 tiles in Appendix F serves as an exact point estimate sensitivity check and does not report bootstrap intervals.

Analysis	Threshold refit?	Variation represented
Primary F_1 and diagnostic headroom	No	Test units; conditional on fitted cut
Primary rankings and paired differences	Yes, each draw	Validation and test units
Cohort without spatial overlap	No	Retained test units; selected cuts fixed
Adapted observers	No intervals	Full split descriptive estimates

B Per-Setting Evaluation Results

Table S1 reports values underlying Table 2 to four decimal places, while Figure S1 compares mean SR performance against matched HR.

When evaluated at the native 1000×1000 resolution without resampling, the same observers achieve F_{val} scores of 0.815, 0.762, 0.823, 0.826, and 0.857 across the five settings, ordered as in Table S1,

Table S1: Setting-specific test split results underlying Table 2. The SR columns report arithmetic means across the six SR pipelines under the native decision (F_{native}), the validation threshold (F_{val}), and the diagnostic headroom ($F^*_{\text{grid}} - F_{\text{val}}$). The HR columns report the same frozen observer evaluated on the HR reference at the matching 1024×1024 resolution; corresponding evaluations at native resolution appear in the text below. Precision indicates the range of test precision across the six SR arms under validation decisions. Grid ceilings use test labels to measure diagnostic headroom.

Task	Model	F_{native}	F_{val}	$F^*_{\text{grid}} - F_{\text{val}}$	HR F_{native}	HR F_{val}	HR F^*_{grid}	HR AUROC	Precision
USBuildings	SegFormer	.0088	.3540	.000089	.8164	.8212	.8212	.9855	.254–.302
USBuildings	Swin	.0029	.2885	.000111	.8200	.8291	.8291	.9870	.191–.252
USBuildings	UNet	.0046	.1945	.001573	.8196	.8233	.8233	.9857	.085–.207
RoadDetection	SegFormer	.0416	.3856	.001055	.8553	.8575	.8577	.9758	.246–.338
RoadDetection	UNet	.0034	.4665	.000187	.8552	.8579	.8580	.9734	.293–.423

with mean AUROC values within images of 0.984, 0.969, 0.986, 0.967, and 0.973. Because the SR pipelines output 1024×1024 images, resampled HR inputs provide the matched reference; evaluations at native resolution reveal how sensitively the reference baseline depends on input dimensions. The largest divergence occurs for Swin on USBuildings, where the reference F_1 at native resolution is 0.067 lower, 0.762 versus 0.829, directly demonstrating why R2 requires identical input dimensions and resampling.

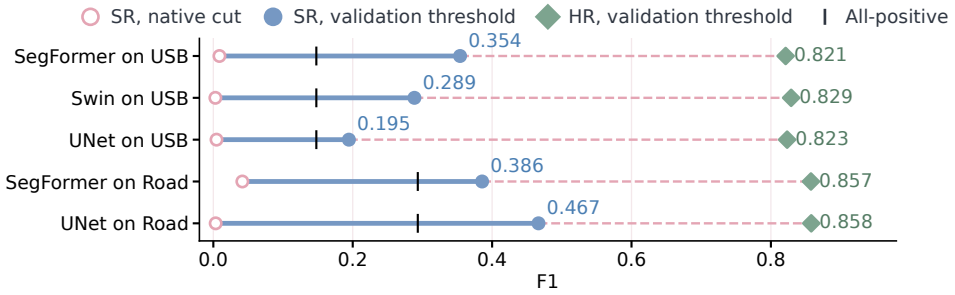

Figure S1: Decision gains leave a substantial gap to matched HR. Mean SR F_1 across the six pipelines under the native cut and the validation threshold alongside matched HR F_1 across all five settings, with the latter two values annotated. HR references pass through the same frozen observer at the matched 1024×1024 input size and threshold policy on identical test tiles. The shift from the native cut to the validation threshold represents the decision gain, while the remaining distance to HR marks the persistent deficit. USB and Road abbreviate USBuildings and RoadDetection.

C Per-Arm Diagnostics

Table S2 catalogs all 40 arms, comprising the 30 SR arms, five bicubic baselines, and five HR references, while Figure S2 maps precision against recall for the SR models.

Across all 30 SR arms, every point estimate under validation decisions exceeds the trivial all-positive baseline. The two narrowest margins occur under UNet on USBuildings, where RGT scores 0.1525 and CAMixerSR scores 0.1570 against the baseline of 0.1478. Because their 95% conditional intervals, $[0.1415, 0.1638]$ and $[0.1462, 0.1680]$, overlap the baseline, superiority remains statistically unresolved in these two cases. The largest diagnostic headroom for any single arm is 0.0075, for CAMixerSR under UNet on USBuildings, followed by 0.0040 for RGT under SegFormer on RoadDetection.

Average precision within images reveals an even more pronounced deficit relative to HR, accompanied by substantial dispersion across arms and tiles. Across the six SR pipelines, mean AP ranges between 0.123 and 0.191 for UNet on USBuildings, the lowest setting, and between 0.272 and 0.442 for SegFormer on RoadDetection, the highest, while the five bicubic arms span 0.147 to 0.208. In contrast, matched HR references at identical input resolutions reach 0.872 to 0.904.

Table S2: Every arm of every setting, test split. The 30 learning-based SR arms, the five bicubic arms, and the five HR reference arms under the frozen observer of each setting. c^* is the validation-selected threshold; F^*_{grid} searches thresholds with test labels and is diagnostic. "AUROC" is the mean within-image AUROC over the 400 test tiles and "< 0.7" the fraction of tiles below 0.7. AP is the tie-aware step average precision and "AP/$\pi \leq 1$" the fraction of tiles whose AP does not exceed that tile's foreground prevalence. HR rows are the reference at the arms' 1024×1024 input size, AP included. * marks arms with no positive prediction at all under the native cut; other displayed zeros are rounded.

Arm	F_{native}	F_{val}	F^*_{grid}	$F^*_{\mathrm{grid}} - F_{\mathrm{val}}$	Prec.	Rec.	AUROC	< 0.7	AP	AP/$\pi \leq 1$	c^*
SegFormer on USBuildings											
ATD	0.0000*	0.3406	0.3406	0.000005	0.254	0.516	0.8061	0.223	0.3092	0.000	-7.248
BiDiffSR	0.0525	0.3605	0.3606	0.000044	0.261	0.581	0.8248	0.140	0.2695	0.000	-4.218
CAMixerSR	0.0000*	0.3465	0.3465	0.000004	0.257	0.533	0.8098	0.230	0.3159	0.000	-7.169
CFAT	0.0000	0.3400	0.3400	0.000045	0.256	0.507	0.8068	0.242	0.3103	0.000	-7.021
RGT	0.0000*	0.3803	0.3805	0.000265	0.302	0.515	0.8243	0.135	0.3449	0.000	-6.734
SRNO	0.0000*	0.3562	0.3564	0.000171	0.268	0.531	0.8044	0.235	0.3133	0.003	-7.159
bicubic	0.0000*	0.2243	0.2244	0.000114	0.136	0.636	0.7107	0.520	0.1941	0.013	-8.665
HR reference	0.8164	0.8212	0.8212	0.000002	0.806	0.837	0.9855	0.000	0.8722	0.000	-0.436
Swin on USBuildings											
ATD	0.0000	0.2638	0.2641	0.000303	0.191	0.429	0.7335	0.435	0.1989	0.003	-6.456
BiDiffSR	0.0171	0.3085	0.3086	0.000071	0.225	0.493	0.7524	0.352	0.1963	0.000	-5.647
CAMixerSR	0.0000	0.2789	0.2790	0.000052	0.202	0.452	0.7465	0.395	0.2048	0.005	-6.064
CFAT	0.0000	0.2774	0.2775	0.000097	0.206	0.424	0.7511	0.405	0.2150	0.000	-6.259
RGT	0.0000	0.2836	0.2837	0.000113	0.209	0.441	0.7591	0.360	0.2193	0.000	-6.025
SRNO	0.0003	0.3188	0.3189	0.000031	0.252	0.435	0.7859	0.290	0.2560	0.000	-5.766
bicubic	0.0001	0.2339	0.2343	0.000410	0.145	0.601	0.7260	0.465	0.1798	0.003	-8.954
HR reference	0.8200	0.8291	0.8291	0.000024	0.815	0.843	0.9870	0.000	0.8814	0.000	-0.599
UNet on USBuildings											
ATD	0.0010	0.1873	0.1877	0.000325	0.137	0.298	0.6197	0.780	0.1571	0.083	-6.479
BiDiffSR	0.0083	0.2394	0.2395	0.000089	0.163	0.449	0.6108	0.865	0.1235	0.058	-6.763
CAMixerSR	0.0021	0.1570	0.1645	0.007488	0.089	0.649	0.5928	0.823	0.1454	0.107	-8.009
CFAT	0.0010	0.1779	0.1785	0.000570	0.155	0.208	0.5936	0.810	0.1502	0.085	-5.938
RGT	0.0002	0.1525	0.1526	0.000130	0.085	0.776	0.5628	0.895	0.1305	0.160	-8.589
SRNO	0.0147	0.2528	0.2537	0.000835	0.207	0.325	0.6717	0.637	0.1914	0.025	-5.533
bicubic	0.0543	0.1480	0.1485	0.000489	0.080	0.999	0.5168	0.860	0.1472	0.393	-12.824
HR reference	0.8196	0.8233	0.8233	0.000023	0.809	0.838	0.9857	0.000	0.8742	0.000	-0.388
SegFormer on RoadDetection											
ATD	0.0026	0.3643	0.3643	0.000031	0.252	0.660	0.6591	0.623	0.2764	0.098	-6.937
BiDiffSR	0.1822	0.4235	0.4235	0.000006	0.312	0.658	0.6692	0.615	0.2721	0.033	-1.561
CAMixerSR	0.0057	0.3689	0.3697	0.000786	0.263	0.617	0.6722	0.585	0.2980	0.090	-6.901
CFAT	0.0040	0.3521	0.3534	0.001308	0.246	0.618	0.6553	0.647	0.2742	0.105	-7.067
RGT	0.0511	0.4365	0.4405	0.003984	0.338	0.617	0.7555	0.325	0.4420	0.020	-5.899
SRNO	0.0041	0.3685	0.3687	0.000217	0.260	0.634	0.6710	0.600	0.2922	0.058	-7.051
bicubic	0.0001	0.2934	0.2934	0.000000	0.172	1.000	0.5040	0.975	0.1712	0.527	-9.779
HR reference	0.8553	0.8575	0.8577	0.000239	0.844	0.871	0.9758	0.005	0.9044	0.000	-0.397
UNet on RoadDetection											
ATD	0.0004	0.4665	0.4666	0.000067	0.390	0.580	0.7089	0.432	0.3352	0.028	-4.381
BiDiffSR	0.0000*	0.4085	0.4088	0.000258	0.293	0.673	0.6652	0.637	0.2341	0.048	-8.045
CAMixerSR	0.0016	0.4841	0.4843	0.000134	0.408	0.596	0.7110	0.443	0.3568	0.037	-4.167
CFAT	0.0003	0.4624	0.4625	0.000037	0.386	0.577	0.6724	0.555	0.3167	0.065	-4.394
RGT	0.0176	0.5172	0.5178	0.000558	0.423	0.666	0.7472	0.278	0.4256	0.035	-4.192
SRNO	0.0003	0.4604	0.4605	0.000066	0.383	0.576	0.6798	0.510	0.3191	0.070	-4.503
bicubic	0.0000	0.3414	0.3415	0.000155	0.261	0.494	0.5352	0.917	0.2085	0.350	-6.370
HR reference	0.8552	0.8579	0.8580	0.000054	0.850	0.866	0.9734	0.005	0.9035	0.003	-0.389

Average precision. Average precision is computed for each test tile from the complete ranking of its 10^6 pixel margins. Tied margins are aggregated into discrete blocks; given distinct threshold values $v_1 > v_2 > \cdots > v_m$ alongside associated recall R_i and precision Pr_i under the decision rule $z \geq v_i$, with initial condition $R_0 = 0$, we calculate $\mathrm{AP} = \sum_i (R_i - R_{i-1})\,\mathrm{Pr}_i$. This formulation evaluates the exact area beneath the precision and recall step curve across block boundaries, omitting both interpolation and right-side maximum smoothing. Because every test tile contains foreground pixels, AP is well-defined across all 400 test tiles in every setting.

D Alternative Decision Rules on Fixed Score Maps

To test whether a richer decision rule could outperform a single global cut, we evaluate four alternative decision rules on the same fixed score maps, three of which operate per image.

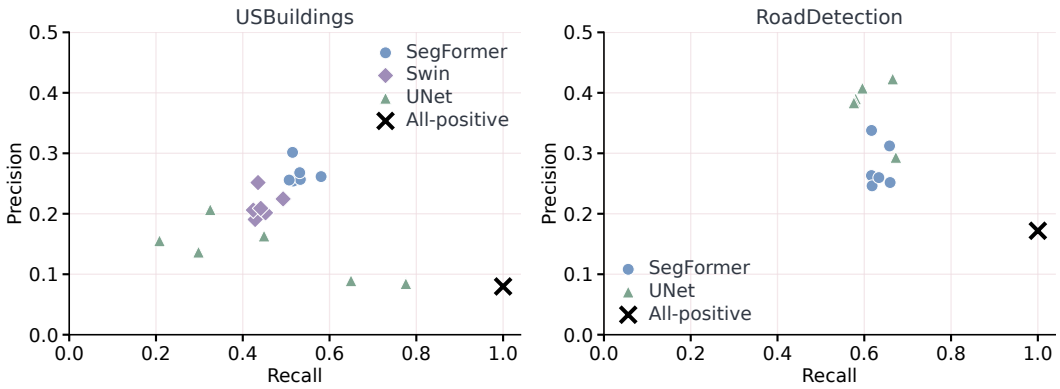

Figure S2: Tradeoffs between precision and recall after threshold selection. Each point represents one of the 30 pairs combining SR pipelines with HR-trained observers under validation decisions, partitioned by task. Marker shapes distinguish the observer, SegFormer, Swin, or UNet. Crosses denote the trivial all-positive rule, which attains a recall of one with precision equal to foreground prevalence.

1. **Fitted per-image top-q.** Each tile labels its highest-scoring fraction q of pixels positive, with a single parameter q shared across tiles and chosen to maximize pooled validation F_1 over the grid $q \in \{0.01, 0.02, \ldots, 1.00\}$, breaking ties toward the smallest q.
2. **Prevalence-matched top-q.** A top-q rule operating per image where q is fixed directly to the validation foreground prevalence π_{val}.
3. **Per-image Otsu thresholding.** Otsu's method (Otsu, 1979) applied to each tile's 256-bin margin histogram, resolving ties toward the lower cut and predicting all-positive if margins are constant.
4. **Rate-matched global threshold.** A single global threshold set at the empirical quantile of pooled validation margins such that the predicted positive rate matches validation prevalence.

For top-q rules, each tile retains its $k = \max(1, \text{round}(qn))$ largest margins, where n denotes the total pixel count per tile, and tied boundary pixels are preserved together. None of these four alternatives achieves a higher mean pooled F_1 across the six pipelines than the primary validation threshold in any setting.

Table S3: Test F_1 for four alternative decision rules applied to the same score maps. Columns report six-pipeline means; top-q is either selected on validation over $q \in \{0.01, \ldots, 1.00\}$ or set to validation prevalence. Otsu operates per image; rate matching sets one global cut. The best alternative mean is 0.010–0.039 below the validation-selected rule in every setting.

			mean pooled F_1			
Task	Model	F_{val}	top-q (fitted)	top-q (π_{val})	Otsu	rate-matched
USBuildings	SegFormer	.3540	.2819	.2739	.2250	.3174
USBuildings	Swin	.2885	.2307	.2206	.2136	.2655
USBuildings	UNet	.1945	.1785	.1657	.1636	.1842
RoadDetection	SegFormer	.3856	.3439	.2883	.3332	.3470
RoadDetection	UNet	.4665	.3818	.3239	.3978	.4271

At the individual arm level, only 8 of the 120 combinations of rules and arms, spanning just 4 of the 30 arms, improve upon the primary validation rule. Across these 8 instances, the median gain is 0.0021 and the maximum is 0.0071.

E Model Rankings and Spatial Sensitivity

Table S4 summarizes ranking stability across the full test split, while Table S5 evaluates stability within each of the two MGRS cells of the USBuildings test split.

For SegFormer on RoadDetection, the difference between BiDiffSR and RGT under validation decisions is -0.0130, with a 95% source-group bootstrap interval of $[-0.0382, +0.0029]$. Because this interval spans zero, the apparent shift in the leading method is not statistically resolved. For USBuildings, shifts in top-ranking pipelines under SegFormer and Swin on the full split do not replicate uniformly across the two MGRS cells. Specifically, the smaller cell of 82 tiles retains BiDiffSR

Table S4: Model-selection stability between native decisions and validation-selected thresholds across the full test split. We compare the ranking of the six learning-based SR pipelines ordered by pooled F_1; τ_b denotes Kendall's tie-corrected rank correlation. The denominator for pairwise inversions includes only pairs strictly orderable under both decision rules; it falls below $\binom{6}{2} = 15$ when arms tie, which occurs under native decisions when multiple pipelines make zero positive predictions. Brackets report 95% percentile bootstrap intervals for τ_b using the validation-refit procedure. The last two columns show the top-ranked pipeline under each rule.

Task	Model	τ_b	95% interval	Inversions	Top-1 native	Top-1 selected
USBuildings	UNet	0.733	[0.600, 1.000]	2/15	SRNO	SRNO
USBuildings	SegFormer	−0.086	[−0.258, 0.346]	5/9	BiDiffSR	RGT
USBuildings	Swin	0.467	[0.200, 0.733]	4/15	BiDiffSR	SRNO
RoadDetection	UNet	1.000	[0.600, 1.000]	0/15	RGT	RGT
RoadDetection	SegFormer	0.733	[0.333, 0.733]	2/15	BiDiffSR	RGT

as the top performer under both decision rules, whereas the larger cell of 318 tiles replicates the shift observed on the full split. For UNet, no top-rank transition occurs on the full split, although the winning pipeline varies across the two cells. Table S5 provides direct comparisons between both individual cells and the pooled full split.

Table S5: Ranking stability within each MGRS cell of USBuildings, test split. The 400 test tiles are partitioned into two MGRS cells containing 82 and 318 tiles. Each row restricts both validation and test splits to that specific cell, refitting the threshold on the cell's own validation tiles, 39 and 121 respectively, rather than transferring a full split threshold to a spatial subset. All uncertainty intervals derive from the validation refit design, which resamples units and refits c^* within every bootstrap draw; because sampling units are individual tiles, these intervals do not account for spatial correlation above the tile level. Point estimates and bootstrap intervals appear in distinct column groups.

		point estimate				separation under resampling			
Split	n	top-1 native	top-1 selected	τ_b	inv./ord.	τ_b 95 % int.	sep.?	top-1 pair	its 95 % int.
UNet on USBuildings									
full	400	SRNO	SRNO	0.733	2/15	[0.600, 1.000]	yes	BiDiffSR − SRNO	[−0.0241, −0.0024]
82-tile cell	82	BiDiffSR	BiDiffSR	0.600	3/15	[0.200, 0.867]	no	BiDiffSR − SRNO	[−0.0071, 0.0347]
318-tile cell	318	SRNO	SRNO	0.733	2/15	[0.600, 1.000]	yes	BiDiffSR − SRNO	[−0.0355, −0.0090]
SegFormer on USBuildings									
full	400	BiDiffSR	RGT	−0.086	5/9†	[−0.258, 0.346]	yes	BiDiffSR − RGT	[−0.0286, −0.0106]
82-tile cell	82	BiDiffSR	BiDiffSR	0.577	0/5†	[0.346, 0.577]	no	BiDiffSR − RGT	[−0.0027, 0.0269]
318-tile cell	318	BiDiffSR	RGT	−0.258	6/9†	[−0.430, 0.346]	yes	RGT − SRNO	[0.0206, 0.0329]
Swin on USBuildings									
full	400	BiDiffSR	SRNO	0.467	4/15	[0.200, 0.733]	yes	BiDiffSR − SRNO	[−0.0173, −0.0032]
82-tile cell	82	BiDiffSR	BiDiffSR	0.467	4/15	[0.200, 0.745]	no	BiDiffSR − SRNO	[−0.0072, 0.0210]
318-tile cell	318	BiDiffSR	SRNO	0.200	6/15	[0.067, 0.447]	yes	BiDiffSR − SRNO	[−0.0240, −0.0079]

† The denominator denotes the number of orderable pairs: pairs tied or undefined under either rule are excluded from both numerator and denominator. Column "sep.?" indicates whether the 95% confidence interval for the difference between the top-1 arm and runner-up excludes zero. The final two columns identify this leading pair and report the confidence interval for their performance difference in the order indicated.

F Threshold Transfer to Held-Out Sources

The RoadDetection test set contains 67 source groups, 39 of which also appear in the validation split. We apply the validation thresholds unchanged to the 48 test tiles from the other 28 groups and measure, on this subset, the retained fraction $(F_{\text{val}} - F_{\text{native}})/(F^*_{\text{grid}} - F_{\text{native}})$, representing the share of recoverable decision gain that the transferred threshold captures. All 12 combinations of the six SR pipelines and two observers retain at least 95.8% of the available gain. Absolute performance F_{val} on these 48 tiles is nevertheless lower than on the 352 shared source tiles for 10 of the 12 combinations.

Cohort without spatial overlap. In addition to shared source groups, validation and test splits may share spatial context through direct geographic overlap. We therefore repeat the threshold transfer analysis on a second subset defined by spatial non-overlap rather than group identifiers. Specifically, we discard any test tile whose spatial footprint intersects that of any validation tile for the same task. This removes 123 of the 400 USBuildings tiles and 7 of the 400 RoadDetection tiles.

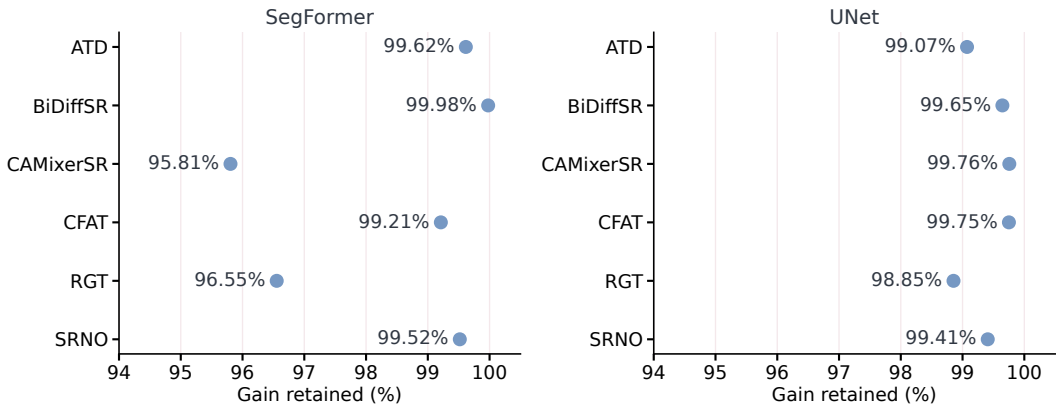

Figure S3: Validation thresholds retain their gains on held-out source groups. Across the 48 RoadDetection test tiles from 28 groups absent from validation, each point shows the retained fraction $(F_{\text{val}} - F_{\text{native}})/(F^*_{\text{grid}} - F_{\text{native}})$ for one SR pipeline and observer under transferred thresholds. Every combination retains at least 95.8% of the available gain, with annotations indicating exact values on this 48-tile subset.

Thresholds are not refitted; each arm retains the cut c^* chosen on the full validation set, restricting only the evaluation cohort.

Table S6: Validation thresholds transferred unchanged to test tiles without spatial footprint overlap, 277/400 USBuildings and 393/400 RoadDetection tiles. Values in F_1 columns are means across the six SR pipelines. HR-relative position (p) is evaluated against each cohort's specific all-positive and matched HR anchors at 1024×1024, with retained HR counts pooled without refitting thresholds. The final column reports the number of pairwise ranking inversions between full and retained cohorts.

Task	Model	F_{val} full	retained	change	p full	p retained	Inv.
USBuildings	SegFormer	0.3540	0.3543	+0.0003	30.6%	34.4%	1/15
USBuildings	Swin	0.2885	0.2870	−0.0015	20.7%	24.7%	0/15
USBuildings	UNet	0.1945	0.1686	−0.0259	6.9%	8.3%	1/15
RoadDetection	SegFormer	0.3856	0.3841	−0.0015	16.4%	16.4%	0/15
RoadDetection	UNet	0.4665	0.4644	−0.0021	30.7%	30.6%	0/15

In four of the five settings, validation-selected F_1 shifts by at most 0.0021, and the leading pipeline remains unchanged across all five settings. The only notable shift occurs for UNet on USBuildings, the lowest-scoring setting, which decreases by 0.0259. Across all three USBuildings settings, the HR-relative position increases because the retained tiles exhibit lower foreground prevalence, thereby lowering the all-positive baseline; for UNet, p rises from 6.9% to 8.3% even as absolute F_1 drops. Although reconstructed images, observer weights, and decision thresholds remain identical, the HR-relative position shifts solely because of cohort composition.

G Controlled Benchmark Analyses

This appendix documents the experimental protocols underlying Section 4.5, covering decision policies on City100, target definitions on BioSR, score transformations on USBuildings, and evaluations on the two published downstream benchmarks.

Decision policy comparisons on City100. City100 (Chen et al., 2019) provides pairs of real low- and high-resolution camera photographs. We evaluate the Nikon subset for fourfold super-resolution, consisting of 10 calibration and 20 test images. Ground truth targets are local maxima of the HR Shi–Tomasi corner response (Shi & Tomasi, 1994) exceeding the image's own 0.94 quantile, extracted via `corner_peaks` with a minimum distance of one pixel. Precision is the fraction of predicted keypoints within two HR pixels of any ground truth target, and recall is the fraction of targets within two HR pixels of any prediction. Keypoint matching is many-to-many, and F_1 is the harmonic mean of precision and recall.

The three reconstruction models share the SmallSR architecture, an eight-block, 64-channel residual network with PixelShuffle and a bicubic skip connection. *l1* minimizes pixelwise L_1 loss; *standard* adds an adversarial loss ($L_1 + 0.02L_{\text{adv}}$) using a patch discriminator with hinge loss; and *zero* uses the same losses as standard, but sets the adversarial gradient to zero for spatial frequencies

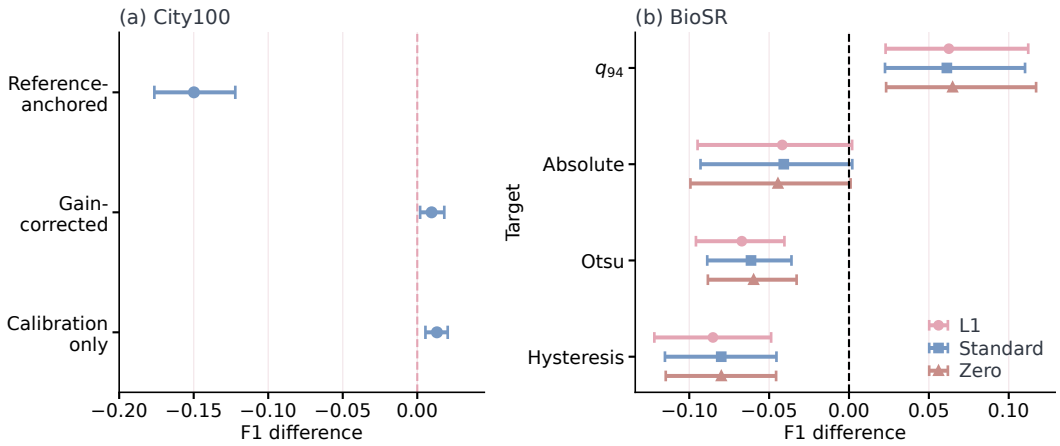

Figure S4: Fixed reconstructions yield policy-dependent and target-dependent comparisons. (a) Difference in F_1 between the zero and standard SmallSR models on City100 under three evaluation policies, comprising a reference-anchored policy, its gain-corrected variant, and a calibration-only policy. (b) Difference in F_1 between the image-level q_{94} rule and the global threshold on BioSR across four target definitions, plotted for the L1, standard, and zero models. Error bars denote paired 95% confidence intervals computed from 10^4 bootstrap resamples of the 20 City100 images or 12 BioSR cells, after averaging three training runs per unit.

$1/6.2338 < f \leq 0.5$ cycles per HR pixel in each RGB channel. This Fourier mask is applied solely during backpropagation; neither forward predictions nor L_1 gradients are filtered. All models are trained for 16 000 iterations across three independent training runs. We score each run independently, average per-image F_1 scores across runs, and then analyze the 20 test images.

We evaluate three decision policies on identical reconstruction pixels:

- INPUT_CAL: Selects a single global threshold per model on calibration images using candidate responses normalized by the bicubic-LR 1st–99th percentile intensity range. This policy relies solely on calibration labels and never accesses test reference images.
- REF: Sets the threshold directly to the 0.94 response quantile of the test reference image; candidate responses use the test HR image's own 1st–99th-percentile intensity stretch. This is a diagnostic stress test that peeks at test ground truth.
- Gain-corrected REF: Divides the REF threshold by a model-specific calibration factor, $g_m = \exp\left(\frac{1}{|\text{cal}|}\sum_{u\in\text{cal}}[\log r_u^{\text{HR}} - \log q_{94}(m,u)]\right)$, matching the model's output response scale to the reference scale. Here r_u^{HR} is the reference response q_{94} threshold and $q_{94}(m,u)$ is the same quantile of model m output response.

We compare the zero and standard models across 10^4 paired bootstrap draws. Switching from REF to INPUT_CAL changes the model contrast by $+0.1630$, with a 95% interval of $[+0.1342, +0.1893]$, and the interaction remains positive, $[+0.1302, +0.1869]$, when thresholds are refitted within each bootstrap draw. The zero-minus-standard contrasts underlying Figure S4a are -0.1498 $[-0.1764, -0.1221]$ under REF, $+0.0096$ $[+0.0019, +0.0181]$ under gain-corrected REF, and $+0.0132$ $[+0.0055, +0.0204]$ under INPUT_CAL.

Target definition comparisons on BioSR. The BioSR microtubule dataset for twofold super-resolution (Qiao et al., 2021) provides 5 calibration and 12 test cells. We train the same SmallSR architecture across three loss configurations using single-channel inputs at scale two, evaluating three independent training runs with the frequency cutoff $1/6.6575 < f \leq 0.5$. Sato curvilinear responses are extracted at scales $1, 1.5, 2$ targeting bright tubular structures. Precision and recall allow a one-pixel Chebyshev tolerance, a 3×3 spatial dilation; F_1 per cell is calculated separately for each run before averaging. Ground truth reference maps are generated by applying a Sato filter (Sato et al., 1998) to HR images, yielding four target definitions:

- q_{94} **target**: Pixels exceeding the reference image's own 0.94 response quantile.
- **Absolute target**: Pixels exceeding a fixed threshold $\theta_{\text{abs}} = 0.0441$, the mean of the five calibration images' q_{94} thresholds, applied uniformly to all images.
- **Otsu target**: Pixels exceeding an image-specific threshold chosen by Otsu's method (Otsu, 1979) on the reference response.

- **Hysteresis target**: Dual-threshold segmentation (Canny, 1986) with high threshold $\theta_{\mathrm{hi}} = \theta_{\mathrm{abs}}$ and low threshold $\theta_{\mathrm{lo}} = \theta_{\mathrm{abs}}/2$.

We evaluate decision rules on the model's output response maps. Here, INPUT_CAL selects a single global threshold per model on calibration cells, whereas OUT_Q94 sets each test image's threshold to its own 0.94 output response quantile, a parameter-free rule structurally aligned with the q_{94} target. Both predictions and fitted decision rules remain frozen across all four target definitions. Paired F_1 contrasts appear in Figure S4b; across the three loss configurations, OUT_Q94 minus INPUT_CAL yields $+0.061$ to $+0.065$ under the q_{94} target, -0.060 to -0.067 under Otsu, and -0.080 to -0.085 under hysteresis, all with intervals excluding zero, alongside -0.041 to -0.045 under the absolute target, with intervals reaching $+0.001$ to $+0.002$.

Score transformation protocols. To evaluate how monotonic score transformations affect decision rules, we test USBuildings margin distributions across three frozen observers and eight input arms, spanning 24 distinct conditions. We evaluate four transformation families:

1. Global gains: $g \in \{0.5, 2, 5, 10\}$;
2. Affine transformations: $(a, b) \in \{(2, 0.01), (0.5, 0.05), (1, 0.02)\}$;
3. Signed powers and logarithms: exponents in $\{0.5, 2\}$ and $\mathrm{sign}(z)\log(1 + \kappa|z|)$ with $\kappa \in \{10, 100\}$;
4. Image-specific random gains: $g_i \sim U[0.5, 2]$.

Candidate threshold grids are recomputed following each transformation. We evaluate two rank-area functionals that integrate maximum precision across recall levels $0, 0.01, \ldots, 1.0$ via the trapezoid rule, computed either by pooling all pixels or by averaging tile-wise areas. Table S7 reports the largest absolute metric shift across all conditions, with the testable sample counts. If the native cut $z > 0$ predicts no positive pixel both before and after a transformation, the metric cannot change, so we drop that case from the testable count.

Table S7: Monotone transformations of the score maps. Each cell gives the largest absolute change in the metric, with the number of testable conditions in parentheses. A condition is untestable when the native cut predicts no positive pixel both before and after the transformation.

Decision functional	gain	affine	power / log	per-image gain
native cut $z > 0$	0 (76/96)	0.006 (58/72)	0 (76/96)	0 (19/24)
calibrated cut, held fixed	0.380 (96/96)	0.356 (72/72)	0.372 (96/96)	0.218 (24/24)
cut re-selected after transform	0 (96/96)	$< 10^{-6}$ (72/72)	$< 10^{-6}$ (96/96)	0.209 (24/24)
quantile rule	0 (96/96)	$< 10^{-6}$ (72/72)	$< 10^{-6}$ (96/96)	$< 10^{-6}$ (24/24)
per-image rank area	0 (96/96)	0 (72/72)	0 (96/96)	0 (24/24)
pooled rank area	0 (96/96)	$< 10^{-6}$ (72/72)	$< 10^{-6}$ (96/96)	0.168 (24/24)

G.1 Published downstream benchmarks: EDTR and SR-COD

Both pipelines are evaluated using their released observers and official evaluation code, without modification. For EDTR (Kim et al., 2025), we evaluate the three arms that emit positive detections on the 5823 PASCAL VOC2012 validation images, scoring micro-F_1 at IoU 0.5 across the confidence grid 0.05–0.95 in steps of 0.01. For SR-COD (Wen et al., 2025), we score the seven released COD models across 2012 COD10K test images, measuring mean per-image IoU over the threshold grid 0.01–0.99 on low-resolution inputs generated by fourfold bicubic downsampling. Thresholds are selected via two-fold cross-fitting on evaluation images, reporting both fold directions.

EDTR. Table S8 lists every confidence cut that the released repository contains, alongside the validation-selected cut and the grid optimum. The selected cut lies between 0.63 and 0.77 in all three arms and in both fold directions, and it captures 87–99% of the headroom available on the grid. The repository's own cuts differ from one another by 0.22–0.35 micro-F_1 within each checkpoint; 0.05 is the score floor inherited from the detection library, while 0.80 and 0.90 appear in the code that renders boxes for figures. The reported metric, mean average precision, consumes every detection above the 0.05 floor and declares no operating point.

SR-COD. Table S9 reports each decision rule per released observer, averaged across the five SR methods. The full 49-cell audit includes seven observers and seven inputs: clean HR, these five SR methods, and a second SwinIR training variant. Three readings follow. The grid optimum sits

Table S8: EDTR decision layer. Micro-F_1 at IoU 0.5 under every confidence cut present in the released repository, under the two-fold validation-selected cut, and at the grid optimum. Fold cuts give the threshold selected on each half. Headroom captured is the validation-selected gain over the default 0.50 as a fraction of the grid optimum's gain over the same default.

	cuts present in the release						selected	
Arm	0.05	0.50	0.80	0.90	F_{val}	F^*_{grid}	fold cuts	captured
Oracle (HR input)	0.4637	0.6691	0.6885	0.6800	0.6895	0.6901	0.77 / 0.74	97%
EDTR, 4 steps	0.2447	0.5748	0.5963	0.5752	0.5977	0.5978	0.73 / 0.75	99%
SR4IR	0.1748	0.4597	0.4739	0.4445	0.4767	0.4791	0.71 / 0.63	87%

between 0.16 and 0.64 depending on the observer, never at the conventional 0.50. The headroom above the default cut is small, at most 0.0047 IoU in any of the 49 cells. And the adaptive cut $\min(2 \cdot \text{mean}, 1)$, which every released COD scorer applies to produce its adaptive F-measure and E-measure, falls below the default 0.50 cut in 43 of the 49 cells, by 0.0123 IoU on average and 0.0289 at worst, more than thirteen times the $+0.0009$ gain available from the best fixed cut on average.

Table S9: SR-COD decision layer. Mean per-image IoU per released COD model, averaged over the five SR methods, at the default cut, at the adaptive cut used by the released scorers, at the two-fold validation-selected threshold, and at the grid optimum. The last column gives the range of the grid-optimal threshold across the five methods. SwinIR uses the DF2K checkpoint.

Observer	IoU @ 0.50	adaptive cut	F_{val}	F^*_{grid}	optimal t
DGNet	0.6148	0.5942	0.6152	0.6152	0.57–0.59
FAPNet	0.5998	0.5829	0.5999	0.6000	0.53–0.58
UR-COD	0.5979	0.5940	0.5981	0.5981	0.30–0.41
SINet-V2	0.5900	0.5625	0.5910	0.5910	0.57–0.64
BGNet	0.5850	0.5770	0.5850	0.5851	0.36–0.52
C2FNet	0.5705	0.5619	0.5705	0.5705	0.43–0.48
SINet	0.5222	0.5248	0.5263	0.5264	0.16–0.18

Method ordering and the matched reference. Averaged across the seven observers, the five SR methods order as ESRGAN at 0.6544 F^w_β, SwinIR at 0.6421, ART at 0.6414, CAT at 0.6401, and EDSR at 0.6396 under the field's continuous metric. Under IoU at the default cut the last two exchange places, 0.5801 for EDSR against 0.5798 for CAT, and the same exchange occurs under the validation-selected threshold, 0.5809 against 0.5807; the exchange appears in five of the seven observers individually, while the leading method is the same under all three orderings. The matched reference is the clean-input arm through the same frozen observers, which reaches 0.6266 under the validation-selected threshold against 0.5923 for the best SR method, a gap of 0.0343 IoU. On EDTR the corresponding gap is 0.6895 against 0.5977, or 0.0918 micro-F_1. Measured from the default cut, the distance to the matched reference exceeds the gain from the decision layer in both pipelines; measured from the 0.05 floor applied in the EDTR release, the decision gap of 0.3530 exceeds the reference gap.

H Threshold Sensitivity under Pipeline-Adapted Observers

Each of the 30 pipeline and observer pairs is evaluated using the adapted checkpoint released by the benchmark (Li et al., 2026). Reconstructions and target labels remain identical under both decision rules, isolating the effect of the threshold. As in the primary audit, we evaluate 160 validation and 400 test tiles, search 2001 candidate empirical quantiles, and break ties in favor of the minimum threshold.

Threshold selection raises the mean F_1 across the six pipelines by 0.1453, 0.1261, and 0.1690 for SegFormer, Swin, and UNet on USBuildings, and by 0.0997 and 0.1055 for SegFormer and UNet on RoadDetection.

Under adapted observers, rank correlation τ_b between native and validation decisions is 0.733 for SegFormer, 0.600 for Swin, and 0.600 for UNet on USBuildings, with $2/15$, $3/15$, and $3/15$ strictly inverted pairs, respectively. On RoadDetection, τ_b reaches 0.867 with $1/15$ inversions for SegFormer and drops to 0.467 with $4/15$ inversions for UNet. Table S10 details every pipeline and observer pair across both observer budgets.

Table S10: Evaluation of all 30 pipeline-observer pairs, comparing the two observer budgets on adjacent rows. H denotes the HR-trained observer and A the pipeline-adapted observer. Both use the same reconstruction, 160 validation tiles, and 400 test tiles. F^*_{grid} is the best test F_1 over the evaluated threshold grid, selected with test labels. Precision and recall use the validation-selected threshold. * marks the cells where the arm predicts no positive pixel at all under the native cut; other displayed zeros are rounded. All entries are full-split point estimates without bootstrap intervals.

Arm	Budget	F_{native}	F_{val}	F^*_{grid}	Prec.	Rec.	c^*
SegFormer on USBuildings							
ATD	H	0.0000*	0.3406	0.3406	0.254	0.516	-7.248
	A	0.4157	0.5516	0.5516	0.471	0.665	-1.030
BiDiffSR	H	0.0525	0.3605	0.3606	0.261	0.581	-4.218
	A	0.2799	0.4802	0.4803	0.396	0.609	-1.266
CAMixerSR	H	0.0000*	0.3465	0.3465	0.257	0.533	-7.169
	A	0.4296	0.5591	0.5591	0.483	0.664	-1.051
CFAT	H	0.0000	0.3400	0.3400	0.256	0.507	-7.021
	A	0.3861	0.5529	0.5529	0.471	0.669	-1.016
RGT	H	0.0000*	0.3803	0.3805	0.302	0.515	-6.734
	A	0.4868	0.5549	0.5558	0.495	0.632	-0.818
SRNO	H	0.0000*	0.3562	0.3564	0.268	0.531	-7.159
	A	0.3776	0.5490	0.5491	0.472	0.657	-1.083
Swin on USBuildings							
ATD	H	0.0000	0.2638	0.2641	0.191	0.429	-6.456
	A	0.4469	0.5751	0.5751	0.497	0.682	-0.959
BiDiffSR	H	0.0171	0.3085	0.3086	0.225	0.493	-5.647
	A	0.3086	0.5046	0.5047	0.420	0.631	-1.206
CAMixerSR	H	0.0000	0.2789	0.2790	0.202	0.452	-6.064
	A	0.4638	0.5806	0.5808	0.507	0.679	-0.876
CFAT	H	0.0000	0.2774	0.2775	0.206	0.424	-6.259
	A	0.4784	0.5745	0.5746	0.498	0.679	-0.900
RGT	H	0.0000	0.2836	0.2837	0.209	0.441	-6.025
	A	0.4974	0.5801	0.5810	0.518	0.659	-0.882
SRNO	H	0.0003	0.3188	0.3189	0.252	0.435	-5.766
	A	0.4333	0.5697	0.5697	0.493	0.675	-0.982

Arm	Budget	F_{native}	F_{val}	F^{*}_{grid}	Prec.	Rec.	c^{*}
UNet on USBuildings							
ATD	H	0.0010	0.1873	0.1877	0.137	0.298	-6.479
	A	0.3691	0.5384	0.5387	0.466	0.637	-1.076
BiDiffSR	H	0.0083	0.2394	0.2395	0.163	0.449	-6.763
	A	0.2452	0.4641	0.4642	0.381	0.594	-1.249
CAMixerSR	H	0.0021	0.1570	0.1645	0.089	0.649	-8.009
	A	0.4076	0.5543	0.5544	0.477	0.661	-1.027
CFAT	H	0.0010	0.1779	0.1785	0.155	0.208	-5.938
	A	0.3398	0.5403	0.5404	0.467	0.641	-1.023
RGT	H	0.0002	0.1525	0.1526	0.085	0.776	-8.589
	A	0.4418	0.5580	0.5586	0.492	0.645	-0.897
SRNO	H	0.0147	0.2528	0.2537	0.207	0.325	-5.533
	A	0.3727	0.5347	0.5348	0.456	0.645	-1.039
SegFormer on RoadDetection							
ATD	H	0.0026	0.3643	0.3643	0.252	0.660	-6.937
	A	0.5821	0.6355	0.6356	0.558	0.737	-0.696
BiDiffSR	H	0.1822	0.4235	0.4235	0.312	0.658	-1.561
	A	0.1620	0.4433	0.4434	0.346	0.618	-1.089
CAMixerSR	H	0.0057	0.3689	0.3697	0.263	0.617	-6.901
	A	0.5539	0.6280	0.6281	0.544	0.742	-0.770
CFAT	H	0.0040	0.3521	0.3534	0.246	0.618	-7.067
	A	0.5838	0.6304	0.6306	0.545	0.747	-0.741
RGT	H	0.0511	0.4365	0.4405	0.338	0.617	-5.899
	A	0.6275	0.6788	0.6788	0.624	0.744	-0.746
SRNO	H	0.0041	0.3685	0.3687	0.260	0.634	-7.051
	A	0.5340	0.6254	0.6255	0.550	0.725	-0.734
UNet on RoadDetection							
ATD	H	0.0004	0.4665	0.4666	0.390	0.580	-4.381
	A	0.5069	0.6110	0.6111	0.527	0.726	-0.948
BiDiffSR	H	0.0000^{*}	0.4085	0.4088	0.293	0.673	-8.045
	A	0.0335	0.3112	0.3113	0.203	0.664	-1.670
CAMixerSR	H	0.0016	0.4841	0.4843	0.408	0.596	-4.167
	A	0.5367	0.6061	0.6062	0.528	0.712	-0.713
CFAT	H	0.0003	0.4624	0.4625	0.386	0.577	-4.394
	A	0.5357	0.6069	0.6072	0.521	0.726	-0.806
RGT	H	0.0176	0.5172	0.5178	0.423	0.666	-4.192
	A	0.6278	0.6588	0.6588	0.595	0.737	-0.724
SRNO	H	0.0003	0.4604	0.4605	0.383	0.576	-4.503
	A	0.5228	0.6024	0.6024	0.518	0.719	-0.802