The morphological barrier in parametric injection-recovery for CNN strong lens finders: evidence from DESI Legacy Survey DR10

A. Author,[1★] B. Author,[2] C. Author[1]
[1] *Institute, Address*
[2] *Institute, Address*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Convolutional neural networks (CNNs) are now widely used to discover strong gravitational lenses in wide-field imaging surveys, but interpreting their yields requires selection functions. A standard calibration is injection-recovery: simulate lensed sources and measure detection completeness as a function of lens and observing conditions. Here we show, using DESI Legacy Imaging Survey DR10 cutouts and an EfficientNetV2-S lens finder, that common parametric injections can be systematically rejected by a high-performing CNN, invalidating naive selection-function interpretation.

At score threshold $p > 0.3$, the model achieves 89.3% recall on 112 spectroscopically confirmed lenses (Tier-A), yet the marginal injection completeness over a multi-dimensional grid is only 3.41% (3755/110000) for Sersic-based lensed sources. We quantify the realism gap directly in the learned representation: a linear probe trained on the 1280-dimensional penultimate features separates real Tier-A lenses from low-impact-parameter injections with AUC 0.996 ± 0.004.

We test the hypothesis that the gap is driven by missing pixel-level noise texture by adding physically correct Poisson shot noise to the injected arcs. Contrary to the texture hypothesis, Poisson noise reduces completeness to 2.37% (2610/110000) at $p > 0.3$ and degrades bright-arc recovery at every magnitude bin. A gain-sweep control (Poisson enabled with gain 10^{12} e^{-}/nmgy) recovers the no-Poisson baseline exactly at every magnitude bin, verifying that the degradation at gain 150 e^{-}/nmgy is physical rather than a code artifact.

These results falsify missing texture as the dominant cause of sim-to-real mismatch in this setting and indicate that morphological realism (source substructure, colour morphology, correlated noise, and PSF fidelity) is the limiting factor for parametric injection-recovery in ground-based data. We provide completeness maps as rigorously characterised conservative lower bounds and propose linear-probe AUC as a practical realism gate for injection pipelines.

Key words: gravitational lensing: strong – methods: data analysis – methods: statistical – surveys – galaxies: structure

1 INTRODUCTION

Strong gravitational lenses enable measurements of galaxy mass profiles, dark matter substructure, and cosmology. Wide-field imaging surveys now deliver many millions of massive galaxies, motivating automated discovery pipelines. CNN-based lens finders can be highly effective, but their scores are not directly interpretable without calibration.

A commonly desired calibration is the *selection function*, the probability that a lens with intrinsic parameters and observing conditions will be detected by a given pipeline. For imaging-based lens finding, selection functions are frequently estimated using *injection-recovery*: simulated lensed sources are injected into real galaxy cutouts, and the detection fraction is measured across parameter space. The validity of this procedure depends on the realism of the injected images. If injections do not occupy the same data manifold as real lenses, injection-recovery can under- or over-estimate detectability in ways that are difficult to diagnose.

This work makes four contributions using verified experiments D01–D05:

(i) A quantitative measurement of the injection realism gap for a CNN lens finder on ground-based survey cutouts: 89.3% Tier-A recall versus 3.41% parametric injection completeness at the same score threshold.

(ii) An experimental falsification of the hypothesis that missing noise texture dominates the gap, using a controlled Poisson shot-noise injection and a gain-sweep control that verifies the code path.

(iii) A feature-space diagnostic of injection realism: a linear probe on penultimate CNN features separates real lenses from injections with AUC 0.996 ± 0.004.

(iv) A rigorously characterised completeness map that is interpretable as a conservative lower bound for the specific injection family used, together with a diagnostic framework to improve injection realism iteratively.

We structure the paper as a falsification ladder: establish the gap, propose a plausible explanation (missing texture), falsify it with controlled experiments, validate the controls, and draw the minimal conclusion supported by the data (a morphological barrier).

[★] E-mail: author@institute.edu

2 DATA

2.1 Survey images and cutouts

We use DESI Legacy Imaging Survey DR10 coadded images in the g, r, and z bands. Each example is a 101×101 pixel cutout at 0.262 arcsec/pixel, centred on a luminous red galaxy (LRG). All analyses are performed on these cutouts after a fixed preprocessing transform (Section 4).

2.2 Label sets

Positives include Tier-A (112 spectroscopically confirmed lenses) and Tier-B (visual candidates). Negatives are LRG-centred cutouts with no known lensing features. The full dataset used for training and validation contains 451,681 cutouts, including 277 Tier-A and 3,079 Tier-B positives.

3 MODEL AND TRAINING

3.1 Architecture

We use EfficientNetV2-S (ImageNet pretrained) with a binary classification head producing a sigmoid score $p \in [0, 1]$.

3.2 Training set-up

Training uses unweighted binary cross-entropy. Augmentations are horizontal and vertical flips. The gen4 checkpoint is obtained by fine-tuning from an earlier checkpoint (gen2 epoch 19) at learning rate 5×10^{-5} for 60 epochs. Data are split 70/30 by HEALPix with zero Tier-A spatial overlap between train and validation. The best validation AUC is 0.9921.

4 PREPROCESSING

All results use the `raw_robust` preprocessing mode: outer-annulus median subtraction and MAD normalisation per band, followed by clipping. The default clip range is $[-10, +10]$. A `clip_range=20` ablation uses $[-20, +20]$. Annulus choices and their effects are documented in Appendix A.

5 INJECTION-RECOVERY FRAMEWORK

5.1 Lens model

Injected lenses use a singular isothermal ellipsoid (SIE) plus external shear. The Einstein radius θ_E spans 0.5–3.0 arcsec in the selection-function grid and is fixed to 1.5 arcsec for the bright-arc experiments. The shear magnitude follows a half-normal with $\sigma = 0.05$. The lens axis ratio is uniform on [0.5, 1.0].

5.2 Source model and priors

Sources are Sersic profiles with optional clumps (60 per cent probability; 1–4 clumps). The selection-function grid uses r-band source magnitude 23–26. The bright-arc experiments sweep lensed apparent magnitudes 18–26. The Sersic index is uniform on [0.5, 4.0], the effective radius is uniform on [0.05, 0.50] arcsec, and the source axis ratio is uniform on [0.3, 1.0]. Colours are drawn from $g - r \sim \mathcal{N}(0.2, 0.25)$ and $r - z \sim \mathcal{N}(0.1, 0.25)$.

5.3 Source position sampling

The source position is parameterised by $\beta_{\rm frac} = \beta/\theta_E$ and sampled area-weighted via $\beta_{\rm frac} = \sqrt{U[\beta_{\rm lo}^2, \beta_{\rm hi}^2]}$. Default $\beta_{\rm frac}$ range is [0.1, 1.0]. Restricted tests use [0.1, 0.55] because high $\beta_{\rm frac}$ cases often produce weakly lensed morphologies and are less relevant for bright-arc controls.

5.4 PSF and observing conditions

The injected arcs are convolved with a per-band Gaussian PSF. The r-band FWHM is read from a manifest and the g and z FWHMs are scaled by 1.05 and 0.94, respectively. The selection-function grid spans 11 θ_E values, 7 PSF bins, and 5 depth bins (22.5–24.5 mag), for 385 cells (220 non-empty, 165 empty). Each non-empty cell contains 500 injections. The grid seed is 1337.

5.5 Why injected arc noise is the missing-texture component

The injection procedure adds simulated arc flux to a real DR10 host cutout. Therefore, the injected images already contain the survey's background and host-galaxy noise and artefacts. The hypothesised missing texture in this setting is primarily the shot noise associated with the added arc flux itself. This motivates injecting Poisson noise into the arc component while leaving the real host and sky pixels unchanged.

6 FALSIFICATION LADDER EXPERIMENTS

6.1 Establishing the realism gap

We compare Tier-A recall (real confirmed lenses) against injection completeness on the multi-dimensional selection-function grid at matched score thresholds. We additionally measure feature-space separability via a linear probe trained on the CNN penultimate features.

6.2 Testing missing texture with Poisson shot noise

We test whether missing pixel-level noise texture is responsible for the realism gap by adding physically correct Poisson shot noise to the injected arcs. For injection image I (nanomaggies) and gain g (electrons per nanomaggy),

$$E = g\,\max(I, 0), \quad E' \sim \mathrm{Poisson}(E), \quad I_{\rm poiss} = I + \frac{E' - E}{g}. \qquad (1)$$

We use $g = 150$ e^{-}/nmgy as an approximate DR10 coadd gain. Zero-flux pixels satisfy $\mathrm{Poisson}(0) = 0$, so no noise is injected into sky-only regions.

6.3 Gain-sweep control

To verify that the Poisson implementation is correct and that any change in detection is not a code artifact, we repeat the bright-arc experiment with Poisson enabled but with gain $g = 10^{12}$ e^{-}/nmgy, making Poisson noise negligible. This should reproduce the no-Poisson baseline exactly.

Table 1. Tier-A recall on 112 spectroscopically confirmed lenses (D05 verified re-evaluation).

Threshold	Recall	$n_{\rm det}/112$	95 per cent Wilson CI
p > 0.3	89.3%	100/112	[82.6%, 94.0%]
p > 0.5	83.9%	94/112	[76.3%, 89.8%]
FPR=1e-3 (p>0.806)	79.5%	89/112	[71.3%, 86.1%]
FPR=1e-4 (p>0.995)	48.2%	54/112	[39.1%, 57.4%]

Table 2. Selection-function grid marginal completeness at $p > 0.3$ (D05).

Condition	Marginal C	Detected	Total
No Poisson (baseline)	3.41%	3,755	110,000
Poisson (gain=150)	2.37%	2,610	110,000

Table 3. Completeness at multiple thresholds (D05).

Threshold	No-Poisson	Poisson	Deficit
p > 0.3	3.41%	2.37%	−1.04pp
p > 0.5	2.75%	1.80%	−0.95pp
p > 0.7	2.26%	1.37%	−0.89pp
FPR=1e-3	1.98%	1.18%	−0.80pp
FPR=1e-4	0.55%	0.25%	−0.30pp

6.4 Bright-arc controlled experiment

We generate bright arcs at fixed $\theta_E = 1.5$ arcsec and measure detection rates versus lensed apparent magnitude in 1-mag bins from 18–26, with N=200 injections per bin and seed=42. Unless otherwise specified, $\beta_{\rm frac} \in [0.1, 0.55]$. This experiment isolates visibility effects and provides a high-signal regime where detection is plausible.

6.5 Linear probe diagnostic

We extract 1280-dimensional penultimate features for real Tier-A lenses and for a controlled injection subset (low $\beta_{\rm frac}$ injections at lensed magnitude 19) and fit a logistic-regression probe with 5-fold cross-validation. We report the mean AUC and its standard deviation across folds, together with Fréchet distances between intermediate feature distributions.

7 RESULTS

7.1 Tier-A recall

7.2 Selection-function completeness is orders of magnitude lower

At $p > 0.3$, injection completeness is 3.41 per cent over the full grid, compared to 89.3 per cent Tier-A recall. This 86-percentage-point discrepancy motivates explicit tests of injection realism rather than assuming injections are a faithful proxy for real lenses.

7.3 Falsification: Poisson shot noise degrades detection

Poisson shot noise at gain 150 e$^-$/nmgy reduces detection in every bright-arc magnitude bin relative to the no-Poisson baseline. This directly contradicts the hypothesis that the realism gap is dominated by missing noise texture. A per-pixel check is consistent: a mag-21 arc spread over about 80 pixels has about 1.05 photoelectrons per pixel at gain 150, so the Poisson standard deviation is comparable to the signal.

7.4 Control: gain sweep recovers the baseline exactly

The gain=10^{12} column in Table 4 matches the baseline exactly at every magnitude bin. With the same Poisson code path but negligible noise, the results are unchanged, verifying that the Poisson implementation is correct and the degradation at gain 150 is physical.

7.5 Completeness structure in parameter space

7.6 Feature-space realism gap

The linear probe AUC of 0.996 indicates near-perfect separability between real confirmed lenses and the injection subset in the CNN representation. The median CNN score differs strongly between the sets (0.995 versus 0.110), consistent with the completeness gap. The Frechet distance increases substantially from early to mid-level features, suggesting that the discriminative mismatch is not limited to shallow pixel noise.

8 DISCUSSION

8.1 Interpretation: a morphological barrier

The key experimental result is that adding physically correct shot noise to injected arcs degrades detectability, and that a gain sweep reproduces the baseline exactly when Poisson variance is made negligible. If missing texture were dominant, adding the missing noise would increase, not decrease, recovery. The minimal conclusion supported by the data is that the injection manifold differs morphologically from the real lens manifold in ways learned by the CNN: source substructure, colour morphology, arc curvature and width distributions, and arc-host interactions not captured by Sersic-plus-clumps injections.

8.2 Why the conclusion is limited and what it is not

We do not claim to have identified a unique morphological feature responsible for the gap. The linear-probe result shows that the CNN representation contains sufficient information to discriminate injections from real lenses, but does not by itself localise which features drive the separation. The Poisson experiment addresses only one class of texture hypothesis: independent shot noise associated with the injected arc flux.

8.3 Completeness as a conservative lower bound

Because injections are demonstrably separable from real lenses in feature space, the measured injection completeness should not be interpreted as an unbiased selection function for the real lens population. It can be interpreted as a conservative lower bound for detectability of the specific parametric injection family under the adopted priors and preprocessing. Practically, this means that any yield extrapolation using these completeness values should be framed as a lower bound unless injection realism is improved and verified with feature-space diagnostics.

Table 4. Bright-arc detection rates at $p > 0.3$ (D05). All: $\theta_E = 1.5$ arcsec, N=200 per mag bin, seed=42, $\beta_{\rm frac} \in [0.1, 0.55]$ unless noted.

Mag bin	Baseline	Poisson (g=150)	clip=20	Poiss+clip20	Unrestricted*	Gain=10^{12}
18–19	17.0%	14.5%	30.5%	31.0%	17.0%	17.0%
19–20	24.5%	18.0%	32.0%	26.5%	21.5%	24.5%
20–21	27.5%	25.5%	37.0%	25.5%	28.0%	27.5%
21–22	35.5%	33.5%	40.5%	24.0%	20.0%	35.5%
22–23	31.0%	29.5%	35.0%	27.5%	17.5%	31.0%
23–24	24.0%	17.5%	14.5%	8.5%	7.0%	24.0%
24–25	8.5%	6.0%	4.5%	1.5%	4.5%	8.5%
25–26	1.0%	1.0%	0.0%	0.0%	0.0%	1.0%

*Unrestricted uses $\beta_{\rm frac} \in [0.1, 1.0]$.

Table 5. Completeness by Einstein radius for the no-Poisson grid at $p > 0.3$ (D05).

θ_E	C(p>0.3)	$n_{\rm det}/n_{\rm inj}$
0.50 arcsec	0.44%	44/10,000
0.75 arcsec	1.22%	122/10,000
1.00 arcsec	2.57%	257/10,000
1.25 arcsec	3.61%	361/10,000
1.50 arcsec	4.33%	433/10,000
1.75 arcsec	4.58%	458/10,000
2.00 arcsec	4.66%	466/10,000
2.25 arcsec	4.44%	444/10,000
2.50 arcsec	4.32%	432/10,000
2.75 arcsec	4.10%	410/10,000
3.00 arcsec	3.28%	328/10,000

Table 6. Completeness by lensed apparent magnitude for the no-Poisson grid at $p > 0.3$ (D05).

Lensed mag	C(p>0.3)	$n_{\rm det}/n_{\rm inj}$
18–20	48.8%	20/41
20–22	20.7%	2,559/12,361
22–24	1.55%	1,082/70,016
24–27	0.34%	94/27,582

Table 7. Linear probe and feature diagnostics (D05).

Metric	Value
Probe AUC (real Tier-A vs low-bf injections)	0.996 ± 0.004 (5-fold CV)
Frechet distance (features_0, early layers)	0.21
Frechet distance (features_3, mid layers)	63.58
Median score: real Tier-A	0.995
Median score: injections (low-bf, mag 19)	0.110
Median score: negatives	1.5e-5

8.4 A realism gate for injection pipelines

The linear-probe AUC provides a practical diagnostic:

(i) Choose a fixed model and preprocessing, and define a real-lens reference set (Tier-A) and an injection set matched to a narrow control slice (for example fixed lensed magnitude and $\beta_{\rm frac}$).
(ii) Extract penultimate features for both sets.
(iii) Train a linear probe with cross-validation to compute AUC.
(iv) Use AUC as a realism gate: improvements to the injection pipeline should reduce AUC while holding controls fixed.

This diagnostic is cheap compared to full retraining and directly targets the learned representation used for discovery.

8.5 Comparison with published work

8.6 Limitations

This paper does not claim that the injection pipeline is realistic. It quantifies and diagnoses the mismatch. Poisson noise is injected only into arc flux, not host or sky, because host and sky noise are inherited from the real cutouts. The PSF model is Gaussian and band scalings are approximate. The Tier-A set is 112 lenses, and the 12 missed lenses at $p < 0.3$ are not yet characterised.

9 FUTURE WORK

Concrete next steps enabled by this diagnostic framework include:

(i) Replacing parametric sources with real galaxy templates (HUDF or GOODS-CANDELS stamps) with realistic colour morphology and clumpy star-forming structure.
(ii) Modelling correlated coadd noise and background systematics beyond independent Poisson sampling.
(iii) Using band-dependent PSF models tied to survey metadata.
(iv) Extending realism tests to cover arc-host blending, dust, and varying source spectral energy distributions.

10 CONCLUSIONS

We have quantified a large injection realism gap for CNN lens finding in DESI DR10: 89.3 per cent Tier-A recall versus 3.41 per cent injection completeness. A linear probe achieves AUC 0.996 separating real lenses from injections in feature space. Adding physically correct Poisson shot noise reduces completeness to 2.37 per cent and degrades bright-arc recovery, falsifying missing texture as the dominant explanation. A gain sweep control at gain 10^{12} reproduces the baseline exactly, validating the Poisson code path. The results indicate a morphological barrier for parametric injection-recovery and motivate feature-space realism diagnostics as an explicit target for injection pipeline development.

ACKNOWLEDGEMENTS

[Placeholder.]

DATA AVAILABILITY

[Placeholder: specify release of code, configuration files, and D05 results.]

Table 8. Comparison with published results.

Paper	Survey	Method	Completeness / Recall	Injection source	Notes
This work	DESI DR10	EfficientNetV2-S	89.3% Tier-A; 3.41% inj.	Sersic + clumps	Probe AUC 0.996
Herle+ (2024)	Euclid sim	ResNet-like	Bias characterized	Sersic	No real lenses
HOLISMOKES XI (2024)	HSC	CNN ensemble	TPR_0 = 10–40%	Real HUDF stamps	Notes Sersic inadequacy
Huang+ (2020)	DECaLS	ResNet	Catalog, no completeness	Sersic	No inj.-recovery map
Jacobs+ (2019)	DES	CNN	∼50% bright arcs	Sersic	Brief test only

APPENDIX A: ANNULUS NORMALISATION CHARACTERISATION

This appendix summarises the effect of the outer-annulus choice used for robust normalisation on 101 × 101 stamps. The original annulus used in earlier code paths was (20,32) pixels. A better-matched outer annulus for 101×101 stamps is (32.5,45) pixels. The median shifts by 0.15 normalised units between the two annuli, but the MAD (robust noise scale) is unchanged (KS $p = 0.648$). Mismatched-annulus scoring shows a 3.6 percentage-point recall drop, which is within the statistical uncertainty for $N = 500$ positives and does not affect the paper's conclusions about injection realism.

This paper has been typeset from a TEX/LATEX file prepared by the author.