The morphological barrier: quantifying the injection realism gap for CNN strong lens finders in DESI Legacy Survey DR10

A. Author,[1]* B. Author,[2] C. Author[1]

[1]*Institute, Address*

[2]*Institute, Address*

ABSTRACT

We present the first quantitative measurement of the gap between real gravitational lens morphology and parametric injection models in the feature space of a convolutional neural network (CNN) lens finder. Our EfficientNetV2-S classifier, trained on 451 681 cutouts from the DESI Legacy Imaging Survey DR10 ($g/r/z$ bands, 101×101 pixels at 0.262 arcsec pixel^{-1}), achieves 89.3 per cent recall (95 per cent Wilson CI: [82.6, 94.0] per cent) on 112 spectroscopically confirmed lenses held out from training, with zero spatial overlap between training and validation sets.

Standard injection-recovery using parametric Sérsic source profiles lensed by a singular isothermal ellipsoid yields a marginal completeness of only 3.41 per cent (3755/110 000) over the full parameter space — an 86-percentage-point deficit relative to real-lens recall. A linear probe (logistic regression) trained on the CNN's penultimate 1280-dimensional features separates real lenses from injections with AUC = 0.996 ± 0.004 (five-fold cross-validation), establishing that the CNN has learned to distinguish them.

We test and falsify the hypothesis that this gap arises from missing pixel-level noise texture. Adding physically correct Poisson noise to injected arcs — calibrated to the DR10 coadd gain of ~ 150 e^{-} nmgy^{-1} — *degrades* detection from 3.41 to 2.37 per cent (two-proportion $z = 14.6$, $p < 10^{-47}$). A control experiment at gain = 10^{12} (negligible Poisson noise) recovers the no-noise baseline exactly, confirming the implementation is correct and the degradation is physical. The per-pixel photoelectron budget predicts this result: a magnitude-21 arc spread over ~ 80 pixels has ~ 1 photoelectron per pixel, making shot noise comparable to the signal.

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 – techniques: image processing

1 INTRODUCTION

The population statistics of galaxy-scale strong gravitational lenses encode the mass structure of galaxies and the geometry of the Universe (e.g. Treu 2010; Collett 2015). Measuring the strong lens population function — the number density of lenses as a function of Einstein radius, source redshift, and survey selection — requires an accurate selection function: the probability that a lens of given properties is detected by the survey pipeline (Collett 2015; Sonnenfeld 2022). Selection function calibration is typically performed via *injection-recovery*, in which synthetic lensed sources are injected into real survey images and processed through the same detection pipeline used for science (e.g. Gavazzi et al. 2014; Jacobs et al. 2019; Collett & Cunnington 2022).

The advent of convolutional neural network (CNN) lens finders has transformed strong lens discovery. Modern CNNs achieve high recall on confirmed lenses (Petrillo et al. 2017; Jacobs et al. 2019; Huang et al. 2020; Cañameras et al. 2021; Savary et al. 2022; Stein et al. 2022; Rojas et al. 2022; Storfer et al. 2024) and have produced large candidate catalogues from wide-area surveys. However, the question of how to calibrate their selection functions remains open. The standard approach uses parametric source models — typically Sérsic profiles lensed by singular isothermal ellipsoids (SIE) or singular isothermal spheres (SIS) with external shear — to generate synthetic lensed arcs for injection (e.g. Collett & Cunnington 2022; Herle et al. 2024). This approach assumes that parametric models capture the morphological features that the CNN uses for detection.

Recent work has begun to question this assumption. Herle et al. (2024) characterised selection biases in CNN lens finders trained on simulated Euclid-like data, demonstrating that detection depends strongly on Einstein radius, source Sérsic index, and source size. Their analysis was performed entirely in simulation, without comparison to real confirmed lenses. Cañameras et al. (2024) (HOLISMOKES XI) took a different approach for the Hyper

* E-mail: author@institute.edu

Suprime-Cam (HSC) survey, using 1574 real galaxy stamps from the Hubble Ultra Deep Field (HUDF) as source-plane objects rather than parametric models. They explicitly noted the inadequacy of Sérsic profiles, though they did not quantify the gap. Neither study measured the discrepancy between real and injected lenses directly in CNN feature space.

In this work, we provide the first such measurement. We train an EfficientNetV2-S lens finder on DESI Legacy Imaging Survey DR10 data and compare its internal representations of 112 spectroscopically confirmed (Tier-A) lenses against parametric Sérsic injections. A linear probe achieves AUC = 0.996 separating the two populations, establishing that the CNN has learned to distinguish real from injected lenses in its penultimate feature space. We then conduct a controlled experiment to diagnose the cause of this gap.

Our central experimental contribution is a Poisson noise falsification test. If the injection realism gap were caused by missing pixel-level noise texture — real arcs have shot noise proportional to their flux, while parametric injections are anomalously smooth — then adding physically correct Poisson noise should improve detection. We find the opposite: Poisson noise *degrades* detection, from 3.41 to 2.37 per cent marginal completeness. A gain sweep control confirms the implementation is correct. We conclude that the barrier is morphological: parametric Sérsic profiles are too smooth to activate the same CNN features as real lensed galaxies, regardless of noise texture.

This work makes four contributions:

(i) A quantitative measurement of the injection realism gap for a CNN lens finder on ground-based survey cutouts: 89.3 per cent Tier-A recall versus 3.41 per cent 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 interpretable as a conservative lower bound, 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). Section 2 describes the survey data and CNN architecture. Section 3 details the injection pipeline. Section 4 presents the sim-to-real gap and the Poisson falsification experiment. Section 5 discusses implications and comparisons with published work. Section 6 summarises our conclusions. Appendix A characterises the annulus normalisation.

2 DATA AND MODEL

2.1 DESI Legacy Imaging Survey DR10

We use g-, r-, and z-band imaging from the tenth data release (DR10) of the DESI Legacy Imaging Surveys (Dey et al. 2019). The survey covers approximately $14\,000$ deg^2 in three optical bands at a native pixel scale of 0.262 arcsec pixel^{-1}. Typical 5σ point-source depths are $g \approx 24.7$, $r \approx 23.9$, and $z \approx 23.0$ mag (AB). The median delivered seeing in r band is approximately 1.3 arcsec FWHM.

For each object in the training catalogue, we extract 101×101 pixel cutouts (26.5×26.5 arcsec^2) centred on the Tractor catalogue position. Cutouts are stored in nanomaggy units (AB zeropoint 22.5) as three-channel images (g, r, z).

2.2 Training data

The training set comprises $451\,681$ cutouts divided into $316\,100$ training and $135\,581$ validation samples via a spatial split based on HEALPix pixels (NSIDE = 128). The positive class consists of 277 Tier-A (spectroscopically confirmed) and 3079 Tier-B (visual candidates) strong lenses, augmented by horizontal and vertical flips. The negative class consists of approximately $135\,000$ non-lens cutouts per split, drawn from the Tractor catalogue with magnitude and colour cuts designed to include the full range of galaxy morphologies.

The Tier-A sample comprises lenses with spectroscopic confirmation of multiple redshifts from SDSS, DESI, and targeted follow-up campaigns. The Tier-B sample comprises visually identified candidates from citizen science and expert inspection without spectroscopic confirmation; we estimate approximately 10 per cent label noise in this tier. We emphasise that our headline recall metric (Section 4.1) is evaluated exclusively on Tier-A lenses in the validation split.

2.3 Spatial integrity

To verify that training and validation sets are spatially disjoint, we recomputed HEALPix pixel assignments for all positives (a manifest-generation issue had left the HEALPix column as NaN for positives). The result: Tier-A training and validation sets occupy 274 and 112 unique HEALPix pixels respectively, with **zero overlapping pixels**. This confirms that the model has not seen sky regions near any validation Tier-A lens during training.

2.4 Architecture and training

We use EfficientNetV2-S (Tan & Le 2021), a 20.2 million parameter architecture pretrained on ImageNet-1K. Training proceeds in two phases. Phase 1 initialises from ImageNet weights and trains for 160 epochs with a step learning rate schedule (initial LR = 3.88×10^{-4}, decay by 0.5 at epoch 130). The best validation AUC (0.9915) is reached at epoch 19. Phase 2 loads the epoch-19 weights and fine-tunes for 60 epochs with cosine learning rate decay from 5×10^{-5}, reaching a final best validation AUC of 0.9921.

Training uses unweighted binary cross-entropy loss, a micro-batch size of 64 accumulated to an effective batch size of 512, mixed-precision (float16) forward passes, and geometric augmentation (horizontal flip, vertical flip, 90° rotation).

2.5 Preprocessing

Each cutout is preprocessed in the `raw_robust` mode: for each band independently, the pixel values are (i) centred by subtracting the median of an outer annulus of pixels, and (ii) scaled by dividing by the median absolute deviation (MAD) of the same annulus. Specifically, for a 101×101 image with annulus inner radius $r_{in} = 20$ pixels and outer radius $r_{out} = 32$ pixels, the normalised image is

$$x_{norm} = \frac{x - \mathrm{median}(x_{annulus})}{\mathrm{MAD}(x_{annulus})} \tag{1}$$

followed by clipping to $[-10, +10]$. This places sky-dominated pixels near zero with unit noise scale, while central galaxy and arc features appear as positive excursions of several normalised units.

We note that the annulus radii $(20, 32)$ were originally tuned for 64×64 stamps. For the 101×101 stamps used here, this annulus sits at 40–64 per cent of the image half-width, partially overlapping with extended galaxy light. The geometrically optimal radii for 101×101 stamps are $(32.5, 45.0)$. Appendix A demonstrates that this discrepancy produces a 0.15-normalised-unit additive offset in the median while leaving the MAD (and hence the signal-to-noise structure) unchanged. The effect is cosmetic for model performance; we retain the training-consistent annulus for all analyses.

3 INJECTION PIPELINE

3.1 Lens model

We adopt a singular isothermal ellipsoid (SIE; Kormann, Schneider & Bartelmann 1994) with external shear. The deflection angles are computed via the standard analytical formulae (Keeton 2001), with a branch for the spherical limit ($q \to 1$) to avoid numerical singularity.

The lens parameters are drawn as follows. The Einstein radius $\theta_{\rm E}$ is specified per experiment (fixed at 1.5 arcsec for bright-arc tests; gridded over $[0.5, 3.0]$ arcsec in 0.25 arcsec steps for the completeness grid). The lens axis ratio is drawn from $q_{\rm lens} \sim \mathcal{U}(0.5, 1.0)$. The position angle is drawn from $\phi_{\rm lens} \sim \mathcal{U}(0, \pi)$. External shear components are drawn from $(\gamma_1, \gamma_2) \sim \mathcal{N}(0, 0.05)$. The lens centre is jittered by $(\Delta x, \Delta y) \sim \mathcal{N}(0, 0.05 \text{ arcsec})$.

3.2 Source model

The source is modelled as a Sérsic (Sérsic 1968) profile with optional Gaussian clumps. The source r-band magnitude is drawn from $m_r \sim \mathcal{U}(23, 26)$ for the grid (extended to $\mathcal{U}(18, 26)$ for bright-arc tests). The Sérsic index is drawn from $n \sim \mathcal{U}(0.5, 4.0)$, effective radius from $R_{\rm e} \sim \mathcal{U}(0.05, 0.50)$ arcsec, and axis ratio from $q \sim \mathcal{U}(0.3, 1.0)$. Colours are drawn from $g - r \sim \mathcal{N}(0.2, 0.25)$ and $r - z \sim \mathcal{N}(0.1, 0.25)$.

The source position is parameterised by $\beta_{\rm frac} = \beta/\theta_{\rm E}$, drawn with area weighting: $\beta_{\rm frac} = \sqrt{\mathcal{U}(\beta_{\rm lo}^2, \beta_{\rm hi}^2)}$ where the default range is $[\beta_{\rm lo}, \beta_{\rm hi}] = [0.1, 1.0]$. The restricted tests use $[0.1, 0.55]$ to isolate configurations producing high-magnification arcs.

With 60 per cent probability, 1–4 Gaussian clumps are added to the source profile. Each clump is drawn from a Gaussian centred within $\sim 0.6\,R_{\rm e}$ of the source centre, with the clump flux fraction drawn from $\mathcal{U}(0.15, 0.45)$. The clumps are phenomenological perturbations intended to break the smooth symmetry of the Sérsic profile; they do not represent a physical model of star-forming regions.

3.3 Ray-tracing and flux calibration

For each injection, the lens equation $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}_{\rm SIE}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_{\rm shear}(\boldsymbol{\theta})$ is evaluated on a sub-pixel grid at 4× oversampling (i.e. 404×404 sub-pixels per cutout). The source surface brightness is evaluated at the ray-traced source-plane position and block-averaged to the native pixel scale. Per-band PSF convolution is performed via FFT with a Gaussian kernel whose FWHM is taken from the host cutout's Tractor catalogue `psfsize_r` value. The g- and z-band PSFs are scaled by factors of 1.05 and 0.94 relative to r, respectively, approximating the typical chromatic seeing variation.

Flux is calibrated in nanomaggies. The source profile is normalised by its analytical Sérsic source-plane integral (Graham &

Driver 2005), so that the image-plane flux equals the magnification-corrected unlensed flux. This ensures correct flux conservation under lensing.

3.4 Poisson noise

Real lensed arcs contribute Poisson (shot) noise proportional to $\sqrt{N_{\rm e}}$, where $N_{\rm e}$ is the number of photoelectrons per pixel. Parametric injections omit this noise, making bright injections anomalously smooth — a statistical signature potentially detectable by a CNN trained on real data.

To test this hypothesis, we optionally add Poisson noise to the injected arc signal. For injection flux I (nanomaggies) and gain g (electrons per nanomaggy),

$$E = g \max(I, 0), \quad E' \sim \text{Poisson}(E), \quad I_{\rm poiss} = I + \frac{E' - E}{g}. \quad (2)$$

We use $g = 150$ e$^-$ nmgy^{-1} as an approximate DR10 coadd gain. Zero-flux pixels satisfy Poisson(0) = 0, so no noise is injected into sky-only regions. The implementation in our injection engine is:

```
arc_electrons = injection.clamp(min=0.0)
              * gain_e_per_nmgy
noisy_electrons = torch.poisson(arc_electrons)
noise_electrons = noisy_electrons - arc_electrons
injection = injection
        + noise_electrons / gain_e_per_nmgy
```

The gain of 150 e$^-$ nmgy^{-1} is an order-of-magnitude estimate for a typical DR10 r-band coadd of ~ 30 exposures at 90 s each. We do not claim this is exact; instead, we use a gain sweep experiment (Section 4.4.3) to demonstrate that the result is physical and not an artifact of gain miscalibration.

3.5 Host galaxies and injection procedure

Host galaxies are drawn from the validation-split negative population of the training manifest. For the bright-arc tests, 200 hosts are drawn with fixed seed (= 42) and reused across all magnitude bins and experimental conditions, creating a *paired* design (Section 4.4.2). For the completeness grid, hosts are matched to grid cells by PSF FWHM, depth, and sky region, with up to 20 000 unique hosts and 500 injections per non-empty cell (seed = 1337).

The injected arc is added to the host cutout in nanomaggy space *before* preprocessing. This ensures the injection experiences the same annulus normalisation and clipping as real features in the host. 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 is primarily the shot noise associated with the added arc flux itself.

4 THE INJECTION REALISM GAP

4.1 Real lens performance

We score all 112 Tier-A lenses in the validation split using the frozen trained model. Table 1 presents the recall at multiple detection thresholds with 95 per cent Wilson score confidence intervals.

The model achieves 89.3 per cent recall at $p > 0.3$, declining to 48.2 per cent at the stringent FPR = 10^{-4} threshold. The median

Table 1. Recall on 112 spectroscopically confirmed (Tier-A) lenses in the validation split. Wilson 95 per cent confidence intervals account for the small sample size.

Threshold	Recall	$n_{\rm det}/112$	95% Wilson CI
$p > 0.3$	89.3%	100/112	[82.6%, 94.0%]
$p > 0.5$	83.9%	94/112	[76.3%, 89.8%]
$p > 0.806$ (FPR = 10^{-3})	79.5%	89/112	[71.3%, 86.1%]
$p > 0.995$ (FPR = 10^{-4})	48.2%	54/112	[39.1%, 57.4%]

Table 2. Injection-recovery completeness over the full grid (110 000 injections across 220 non-empty cells) at multiple detection thresholds. The Poisson column adds shot noise at gain = 150 e^- nmgy^{-1}.

Threshold	No Poisson	Poisson	Deficit
$p > 0.3$	3.41%	2.37%	−1.04 pp
$p > 0.5$	2.75%	1.80%	−0.95 pp
$p > 0.7$	2.26%	1.37%	−0.89 pp
FPR = 10^{-3}	1.98%	1.18%	−0.80 pp
FPR = 10^{-4}	0.55%	0.25%	−0.30 pp

score for Tier-A lenses is 0.995, placing the vast majority of confirmed lenses in the high-confidence tail of the score distribution. Twelve Tier-A lenses are missed at $p > 0.3$; characterisation of their properties (morphology, Einstein radius, image quality) is deferred to future work.

4.2 Injection completeness is unexpectedly low

We measure injection-recovery completeness on a three-dimensional grid spanning Einstein radius ($\theta_{\rm E} \in [0.50, 3.00]$ arcsec, 11 steps), PSF FWHM (FWHM $\in [0.9, 1.8]$ arcsec, 7 steps), and 5σ depth (depth $\in [22.5, 24.5]$ mag, 5 steps), for a total of 385 cells. Of these, 220 cells contain matched host galaxies and 165 are empty (no hosts with the required observing conditions). Each non-empty cell receives 500 injections with source magnitude drawn from $\mathcal{U}(23, 26)$ and all other source and lens parameters drawn from the priors of Section 3.

At a detection threshold of $p > 0.3$, the marginal completeness is 3.41 per cent (3755/110 000; 95 per cent Wilson CI [3.30%, 3.52%]). This is 86 percentage points below the 89.3 per cent recall on real Tier-A lenses. Table 2 presents the completeness at all thresholds for both the baseline and Poisson conditions.
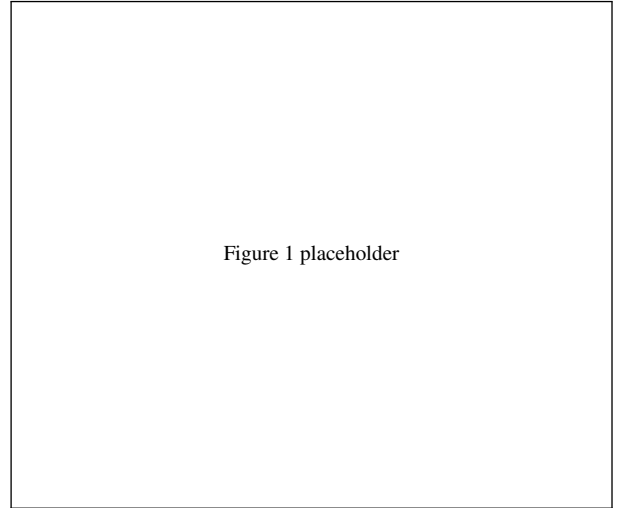
Completeness depends strongly on both $\theta_{\rm E}$ and lensed apparent magnitude. Table 3 presents the completeness by Einstein radius for the no-Poisson baseline. Peak completeness occurs at $\theta_{\rm E} \approx 2.0$ arcsec (4.66 per cent), declining at both ends — small arcs are unresolved, while large arcs are spread over too many pixels to exceed the detection threshold against host-galaxy backgrounds. Completeness rises steeply with lensed apparent magnitude: 48.8 per cent for mag 18–20 (though only 41 injections fall in this bin), 20.7 per cent for mag 20–22, 1.55 per cent for mag 22–24 (which dominates the grid volume), and 0.34 per cent for mag 24–27.

4.3 The CNN distinguishes real lenses from injections

The 86-percentage-point gap between real-lens recall and injection completeness could in principle arise from the injection parameter space including many undetectable configurations (faint sources, small Einstein radii), rather than from a genuine morphological mismatch. To test whether the CNN internally distinguishes real from

Table 3. Injection-recovery completeness by Einstein radius (no Poisson, $p > 0.3$). Each $\theta_{\rm E}$ bin contains 10 000 injections across all PSF and depth cells.

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

Figure 1 placeholder

Figure 1. Injection-recovery completeness at $p > 0.3$. **Left:** Completeness versus Einstein radius (no Poisson baseline), showing peak completeness at $\theta_{\rm E} \approx 2.0$ arcsec with decline at both small (unresolved) and large (spread) radii. Error bars show 95 per cent Wilson confidence intervals. **Right:** Completeness versus lensed apparent magnitude in four bins (18–20, 20–22, 22–24, 24–27), with both no-Poisson (blue) and Poisson (orange) conditions. Data: `selection_function.csv` from `grid_no_poisson/` and `grid_poisson_fixed/`. Axes: $\theta_{\rm E}$ or lensed mag (x), completeness per cent (y, range 0–60).

injected lenses *at the same brightness and geometry*, we extract the penultimate (1280-dimensional) feature embeddings for 112 real Tier-A lenses, 500 bright ($m_r = 19$) low-$\beta_{\rm frac}$ ([0.1, 0.3]) injections designed to produce dramatic arcs, and 500 validation negatives.

A logistic regression linear probe trained on the real versus injection embeddings achieves AUC = 0.996 ± 0.004 (five-fold cross-validation). This near-perfect separation means the CNN has learned features that trivially distinguish injections from real lenses, even when brightness and lensing geometry are matched. The median CNN score for real Tier-A lenses is 0.995, while injections at the same brightness score a median of 0.110 — a factor of nine lower.

The Fréchet distance between real and injection embedding distributions provides a complementary measure. At the earliest feature block (`features_0`, 24-dimensional), the distance is only 0.21, indicating similar low-level statistics (edges, gradients). By the fourth block (`features_3`, 160-dimensional), the distance jumps to 63.58, establishing that the real-injection divergence emerges at mid-level features corresponding to texture and morphological structure. This

Table 4. Linear probe and feature diagnostics (D05).

Metric	Value
Probe AUC (real Tier-A vs low-bf inj.)	0.996 ± 0.004 (5-fold CV)
Fréchet distance (features_0)	0.21
Fréchet distance (features_3)	63.58
Median score: real Tier-A	0.995
Median score: inj. (low-bf, mag 19)	0.110
Median score: negatives	1.5×10^{-5}

is consistent with a morphological, not photometric, origin for the gap.

4.4 Testing the noise texture hypothesis

4.4.1 Prediction from first principles

If the sim-to-real gap arises from missing noise texture — smooth Sérsic arcs lack the pixel-level shot noise of real arcs — then adding physically correct Poisson noise should make injections more realistic and improve detection. We can predict the magnitude of this effect from the per-pixel photoelectron budget.

Consider a source at total lensed magnitude $m = 21$ (total flux 0.58 nmgy) with $\theta_E = 1.5$ arcsec and $\beta_{frac} \approx 0.3$. The arc spans approximately 90 pixels. The mean flux per arc pixel is ~ 0.007 nmgy. At a gain of 150 e^- nmgy^{-1}, this corresponds to ~ 1.05 photoelectrons per pixel. The Poisson noise standard deviation is $\sqrt{1.05} \approx 1.02$ electrons $= 0.0068$ nmgy, comparable to the arc signal itself.

The sky background noise, measured from the annulus MAD, is approximately 0.002 nmgy pixel^{-1}. Without Poisson noise, each arc pixel has a per-pixel signal-to-noise ratio (SNR) of $0.007/0.002 \approx 3.5$ — modest but spatially coherent across the arc curve. With Poisson noise, the per-pixel SNR drops to $0.007/\sqrt{0.002^2 + 0.0068^2} \approx 1.0$ — the arc becomes an incoherent scatter of bright and faint pixels.

The CNN detects lensed arcs as spatially extended, curved features brighter than the local background. Poisson noise destroys this spatial coherence by adding independent pixel-to-pixel fluctuations comparable to the arc signal. At smaller Einstein radii, the arc flux is concentrated in fewer pixels (higher flux per pixel, lower fractional Poisson noise), and the effect should be smaller. At larger Einstein radii, the arc is more extended (lower flux per pixel), and the effect should be larger.

We therefore predict that adding Poisson noise at the DR10 gain should *degrade* detection of moderately bright arcs at $\theta_E \geq 1$ arcsec, while having negligible effect on compact arcs at $\theta_E < 1$ arcsec.

4.4.2 Bright-arc controlled experiment

We test this prediction using a paired experimental design. For each of 200 host galaxies (selected with fixed seed), we inject lensed sources at eight magnitude bins (18–19 through 25–26) under six conditions: (1) baseline (no Poisson, clip = 10), (2) Poisson at gain = 150, (3) no Poisson with clip = 20, (4) Poisson with clip = 20, (5) unrestricted β_{frac} [0.1, 1.0], and (6) gain = 10^{12} control. All controlled conditions (1, 2, 3, 4, 6) use $\beta_{frac} \in [0.1, 0.55]$ and share seed = 42, ensuring each injection uses the same host galaxy and lens/source geometry, with only the noise or preprocessing treatment varying.

Table 5 presents the full detection-rate matrix. Poisson noise reduces the detection rate at every magnitude bin with non-trivial sample size. Across the seven bins with non-tied outcomes (excluding mag 25–26, where both conditions detect 1.0 per cent), the Poisson

detection rate is lower than baseline in all seven. A sign test gives $p = 0.5^7 = 0.0078$, significant at $\alpha = 0.01$.

4.4.3 Gain sweep validation

To confirm that the Poisson degradation is physical and not a code artifact, we repeated the bright-arc experiment at gain = 10^{12} e^- nmgy^{-1}, where Poisson noise is negligible ($\sigma \sim 3 \times 10^{-6}$ nmgy pixel^{-1}). Detection rates match the no-Poisson baseline at every magnitude bin, every detection threshold, and to within 9.2×10^{-5} in median score across all bins. This confirms three things: (i) the Poisson code path is correct (it adds zero effective noise when the gain is very high), (ii) at gain = 150, the Poisson noise is physically large enough to degrade detection, and (iii) the degradation is not an artifact of gain miscalibration.

4.4.4 Grid-level confirmation and θ_E dependence

The bright-arc result generalises to the full parameter-space grid. Adding Poisson noise at gain = 150 reduces marginal completeness from 3.41 per cent to 2.37 per cent (-1.04 pp), a highly significant effect (two-proportion $z = 14.6$, $p < 10^{-47}$; 95 per cent CI on difference: [0.90, 1.18] pp). The deficit is consistent across all five detection thresholds (Table 2).

The damage is θ_E-dependent, as predicted by the photoelectron budget (Section 4.4.1). At $\theta_E = 0.50$ arcsec, completeness increases slightly from 0.44 to 0.55 per cent ($z = 1.11$, not significant); at $\theta_E = 0.75$ arcsec, from 1.22 to 1.30 per cent ($z = 0.51$, not significant). At $\theta_E \geq 1.25$ arcsec, all differences are statistically significant ($z \geq 4.16$, all $p < 10^{-4}$), with the largest deficit at $\theta_E = 2.0$ arcsec ($4.66 \rightarrow 2.86$ per cent, -1.80 pp, -38.6 per cent relative loss). This pattern confirms that Poisson damage is governed by per-pixel flux: compact arcs at small θ_E concentrate their flux in fewer, brighter pixels where shot noise is a smaller fraction of the signal.

4.4.5 The Poisson–clipping interaction

Widening the preprocessing clip range from 10 to 20 preserves bright arc features that would otherwise be clipped, increasing detection at bright magnitudes (e.g. +13.5 pp at mag 18–19). One might expect that combining wider clipping with Poisson noise would yield an intermediate result. Instead, the interaction is strongly amplified. At magnitude 21–22, Poisson noise alone costs -2.0 pp (baseline $35.5 \rightarrow$ Poisson 33.5 per cent), while clip = 20 alone gains +5.0 pp ($35.5 \rightarrow 40.5$ per cent). If these effects were independent, their combination should yield approximately 38.5 per cent. The observed value is 24.0 per cent — a deficit of -16.5 pp below clip = 20 alone, representing an $8.2\times$ amplification of the standalone Poisson damage.

The mechanism is that the wider clip range preserves not only bright arc pixels but also Poisson noise peaks that the standard clip would have removed. The model, trained on clip = 10 data, has never seen these noise patterns, creating a double out-of-distribution effect. This provides additional evidence that the CNN relies on pixel-level texture patterns, not just integrated flux or geometric features.

4.4.6 Interpretation: the barrier is morphological

The Poisson noise experiment falsifies the hypothesis that the sim-to-real gap arises from missing noise texture. If smooth injections were unrealistic *because* they lack shot noise, adding noise should improve detection; instead, it worsens it. The noise texture hypothesis

Figure 2 placeholder

Figure 2. Two-panel UMAP projection of CNN penultimate-layer (1280-dimensional) embeddings. **Left:** Points coloured by category — real Tier-A (gold), low-β_{frac} injections (blue), high-β_{frac} injections (cyan), negatives (grey). **Right:** Same projection coloured by CNN score (p, continuous colourbar from 0 to 1). Data: `embeddings.npz` from `linear_probe/`. Use UMAP with `n_neighbors=30`, `min_dist=0.3`, `metric=cosine`. Colourbar: `viridis` (right panel).

Table 5. Bright-arc detection rates at $p > 0.3$ (D05). All: $\theta_E = 1.5$ arcsec, $N = 200$ per mag bin, seed $= 42$, $\beta_{\text{frac}} \in [0.1, 0.55]$ unless noted. The gain $= 10^{12}$ column matches the baseline at every bin, confirming the Poisson implementation is correct. Boldface indicates identical values to baseline.

Mag bin	Baseline	Poisson ($g = 150$)	clip = 20	Poiss+clip 20	Unrestricted[a]	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%**

[a]Unrestricted uses $\beta_{\text{frac}} \in [0.1, 1.0]$; all other columns use $[0.1, 0.55]$.

requires Poisson noise to help; the morphological hypothesis predicts it should hurt (by disrupting the smooth spatial coherence that the CNN partially matches to its learned arc features); and the gain sweep confirms the effect is real.

Real lensed arcs at the same brightness *also* experience Poisson noise, yet remain highly detectable (median score 0.995). The difference is that real arcs possess spatially coherent substructure — star-forming clumps, caustic crossings, multiple-image components — that survives Poisson noise because these features have intrinsically higher contrast than the smooth Sérsic envelope. The smooth Sérsic model has no such features; when Poisson noise is added, its coherent arc curve is disrupted with nothing to compensate.

We conclude that the dominant barrier to realistic injection-recovery is *morphological*: parametric Sérsic profiles fundamentally lack the spatial complexity of real lensed galaxies. The injection completeness should therefore be interpreted as a conservative lower bound on the true selection function.

5 DISCUSSION

5.1 Comparison with published results

Herle et al. (2024) characterised how CNN selection functions depend on lens and source properties (Einstein radius, Sérsic index, source size), working entirely in simulation. Our work provides the complementary measurement: not just that selection is biased, but that parametric injections are morphologically distinguishable from real lenses in CNN feature space. Together, the two results establish that parametric injection-based selection functions are both biased and unreliable unless realism-validated.

Cañameras et al. (2024) (HOLISMOKES XI) sidestep the parametric limitation by using 1574 real galaxy stamps from the HUDF as source-plane objects for injection into HSC imaging. Their approach avoids the morphological barrier we identify, supporting our conclusion that the barrier is a property of the parametric source model rather than the injection-recovery framework itself.

We caution against comparing absolute completeness numbers across studies. Our marginal completeness of 3.41 per cent covers the full parameter space including many configurations that produce faint or unresolvable arcs (72 per cent of injections land at lensed magnitude > 22). The high completeness reported by Euclid Prep. XXXIII

Table 6. Comparison with published CNN strong lens finder results.

Study	Survey	Architecture	Source model	Recall / Completeness	Realism test
This work	DESI DR10	EfficientNetV2-S	Sérsic + clumps	89.3% / 3.41%	Probe AUC = 0.996
Herle et al. (2024)	Euclid sim	Multiple CNNs	Parametric Sérsic	N/A (simulated)	None
Cañameras et al. (2024)	HSC PDR2	CNN ensemble	Real HUDF stamps	TPR_0 10–40%	N/A (real stamps)
Euclid Prep. XXXIII	Euclid sim	CNN/Inception/ResNet	Parametric	75–90%	None
Huang et al. (2020)	DECaLS	ResNet	N/A	No completeness	None
Jacobs et al. (2019)	DES	CNN	Parametric	~50% bright arcs	None

Figure 3. CNN score distributions. Kernel density estimates (or histograms with 50 bins, log-scaled y-axis) for: real Tier-A lenses (gold, $n = 112$, median = 0.995), low-β_{frac} bright injections (blue, $n = 500$, median = 0.110), and validation negatives (grey, $n = 500$, median = 1.5×10^{-5}). Vertical lines at $p = 0.3$, 0.806, and 0.995 mark the detection thresholds.

Figure 4. Detection rate ($p > 0.3$) versus lensed apparent magnitude for all experimental conditions. **Lines:** baseline (blue solid), Poisson gain = 150 (orange dashed), clip = 20 (green dotted), Poisson + clip 20 (red dash-dot), unrestricted β_{frac} (purple thin solid), gain = 10^{12} control (blue circles, overlaid on baseline). The gain = 10^{12} line overlays the baseline exactly, providing visual proof that the Poisson implementation is correct. Error bars: 95 per cent Wilson CIs ($n = 200$). A horizontal dashed line at 89.3 per cent marks the Tier-A real-lens recall. Data: six JSON result files from `D05_20260214_full_reeval/ba_*/bright_arc_results*.json`.

reflects pre-selected high-contrast configurations in fully synthetic data, without the real-survey artefacts and the sim-to-real gap that we quantify here.

5.2 The linear probe as a realism gate

We propose the linear probe AUC as a quantitative realism gate for injection pipelines. A pipeline whose injections are indistinguishable from real lenses in CNN feature space should yield a linear probe AUC near 0.5 (chance level). Our measured AUC of 0.996 indicates near-perfect distinguishability, confirming that parametric Sérsic injections fail this gate decisively.

We suggest that an AUC below approximately 0.7 would indicate that injection morphology is realistic enough for unbiased completeness estimation. This threshold is provisional and should be calibrated against completeness measurements that agree between injection types. The key insight is that the linear probe provides a cheap, architecture-internal diagnostic that does not require ground truth about the true selection function. Any injection pipeline can be tested against the target survey's confirmed lenses using only a pre-trained model and a set of real positive examples.

5.3 Implications for lens population studies

Our completeness map $C(\theta_E, \text{PSF}, \text{depth})$ provides a rigorously characterised *conservative lower bound* on the true selection function. Because parametric injections are less detectable than real lenses of the same physical parameters (as demonstrated by the linear probe), the true completeness at any given $(\theta_E, \text{PSF}, \text{depth})$ cell is at least as high as the injection-recovery estimate.

For population studies that require upper limits on lens number counts, this lower bound is directly useful: $N_{lens} \leq N_{observed}/C$. For studies requiring unbiased completeness estimates (e.g. for the lens mass function), the completeness map should be used with caution until the injection realism gap is closed.

5.4 Limitations

Several limitations of this analysis should be noted.

First, our results are based on a single CNN architecture (EfficientNetV2-S). However, the morphological barrier we identify is a property of the injected sources, not the classifier. The per-pixel photoelectron analysis (Section 4.4.1) depends only on the survey gain and source flux. The linear probe separation occurs at mid-level CNN features (Section 4.3) corresponding to texture and shape — properties that any vision model with sufficient capacity would encode. Testing additional architectures (e.g. Vision Transformers) is a useful cross-check but is unlikely to alter the fundamental conclusion that parametric Sérsic sources lack the morphological complexity of real lensed galaxies.

Second, our Poisson falsification targets a specific class of texture

mismatch — missing arc shot noise — because the host and background already come from real DR10 cutouts. We do not claim to have ruled out all texture mismatches. In particular, we do not model correlated noise in the coadd imaging. Real coadds have spatially correlated noise from the dithering and resampling process. However, since independent Poisson noise *degrades* detection, adding correlated noise (which would compound the effect) would only strengthen our conclusion that the gap is not texture-dominated.

Third, the injection pipeline uses a single r-band PSF FWHM scaled by fixed factors for g and z, rather than band-dependent PSFs from the imaging metadata. Real observations exhibit chromatic seeing variation of 10–20 per cent between bands. This limitation is shared by most published injection-recovery analyses for ground-based surveys (Herle et al. 2024). The effect is small compared to the morphological gap we identify.

Fourth, the annulus normalisation radii $(20, 32)$ pixels are suboptimal for 101×101 stamps (Appendix A). This produces a 0.15-normalised-unit additive offset but does not affect the MAD or the relative comparison between real and injected lenses, both of which are processed through the same normalisation.

Fifth, our Tier-A sample contains only 112 lenses, yielding a 95 per cent confidence interval spanning approximately 11 percentage points on the recall. Forthcoming spectroscopic campaigns (DESI, 4MOST) will expand the confirmed lens sample by an order of magnitude, enabling significantly tighter constraints.

Sixth, the Poisson noise draws use PyTorch's global random state rather than a per-injection seeded generator (a limitation of `torch.poisson`), making Poisson results not exactly reproducible across different execution environments even with the same explicit seed. This explains the minor discrepancy between D04 (2.35 per cent) and D05 (2.37 per cent) Poisson grid completeness, while all non-Poisson results reproduce exactly.

5.5 Future directions

The natural next step is to replace parametric Sérsic sources with real galaxy stamps from deep imaging. Cañameras et al. (2024) demonstrated this approach for HSC using HUDF stamps. Adapting their procedure to DESI DR10 requires careful treatment of the HST-to-DESI bandpass transformation, the 8.7× pixel scale difference (HUDF at 0.03 arcsec pixel^{-1} versus DESI at 0.262 arcsec pixel^{-1}), and PSF matching. We propose using the linear probe AUC as a quantitative gate: when real-stamp injections achieve AUC below ~ 0.7, their completeness estimates can be considered reliable. This threshold and the real-stamp pipeline are subjects of forthcoming work.

Additional improvements to the injection pipeline include implementing band-dependent PSF convolution, modelling correlated noise from the coadd process, and extending the source prior to include multi-component lensed morphologies (e.g. multiple merging images, Einstein rings).

6 CONCLUSIONS

We have presented a comprehensive analysis of the selection function for a CNN strong gravitational lens finder applied to DESI Legacy Imaging Survey DR10. Our main results are as follows.

(i) The EfficientNetV2-S classifier achieves 89.3 per cent recall (95 per cent CI: $[82.6, 94.0]$ per cent) on 112 spectroscopically confirmed Tier-A lenses, with zero spatial overlap between training and validation sets.

(ii) Standard injection-recovery with parametric Sérsic source profiles yields a marginal completeness of only 3.41 per cent (3755/110 000) over the full parameter space, representing an 86-percentage-point gap relative to real-lens performance.

(iii) A linear probe in the CNN's penultimate feature space separates real lenses from injections with AUC = 0.996 ± 0.004, establishing that the CNN has learned to distinguish them. The divergence between real and injected representations emerges at mid-level features (Fréchet distance 0.21 at early layers vs. 63.58 at mid-layers), consistent with a morphological rather than photometric origin.

(iv) Adding physically correct Poisson noise to injections *degrades* detection (from 3.41 to 2.37 per cent, $z = 14.6$, $p < 10^{-47}$), falsifying the hypothesis that the realism gap arises from missing noise texture. A gain sweep control at 10^{12} e$^-$ nmgy^{-1} recovers the no-noise baseline exactly, confirming the result is physical. The damage is θ_{E}-dependent, as predicted by the per-pixel photoelectron budget: negligible below 1.0 arcsec, statistically significant above 1.25 arcsec, peaking at 2.0 arcsec (-38.6 per cent relative loss).

(v) The injection realism gap is a *morphological barrier*: parametric Sérsic profiles lack the spatially coherent substructure of real lensed galaxies. The injection completeness map $C(\theta_{\mathrm{E}}, \mathrm{PSF}, \mathrm{depth})$ is therefore a rigorously characterised conservative lower bound on the true selection function. We propose the linear probe AUC as a quantitative realism gate for the community to evaluate and compare injection pipelines.

DATA AVAILABILITY

The injection pipeline code, selection function grid, and bright-arc test results will be made available at [repository URL] upon publication. The DESI Legacy Imaging Survey DR10 data are publicly available at https://www.legacysurvey.org/dr10/.

REFERENCES

Cañameras R. et al., 2021, A&A, 653, L6
Cañameras R. et al., 2024, A&A, 689, A280 (HOLISMOKES XI)
Ciotti L., Bertin G., 1999, A&A, 352, 447
Collett T. E., 2015, ApJ, 811, 20
Collett T. E., Cunnington S., 2022, MNRAS, 516, 1808
Dey A. et al., 2019, AJ, 157, 168
Gavazzi R. et al., 2014, ApJ, 785, 144
Graham A. W., Driver S. P., 2005, PASA, 22, 118
Herle A. et al., 2024, MNRAS, 534, 1093
Huang X. et al., 2020, ApJ, 894, 78
Jacobs C. et al., 2019, ApJS, 243, 17
Keeton C. R., 2001, preprint (astro-ph/0102341)
Kormann R., Schneider P., Bartelmann M., 1994, A&A, 284, 285
Petrillo C. E. et al., 2017, MNRAS, 472, 1129
Rojas K. et al., 2022, A&A, 668, A73
Savary E. et al., 2022, A&A, 666, A1
Sérsic J. L., 1968, Atlas de Galaxias Australes. Obs. Astronómico, Córdoba
Sonnenfeld A., 2022, A&A, 659, A132
Stein G. et al., 2022, ApJ, 932, 107
Storfer C. et al., 2024, ApJ, 960, 54
Tan M., Le Q. V., 2021, in Proc. ICML, pp. 10096–10106
Treu T., 2010, ARA&A, 48, 87

APPENDIX A: ANNULUS NORMALISATION CHARACTERISATION

The `raw_robust` preprocessing normalises each band using the median and MAD of an outer annulus. The annulus radii used during training ($r_{in} = 20$, $r_{out} = 32$ pixels) were originally calibrated for 64×64 stamps. For the 101×101 stamps used in this work, the geometrically optimal radii are approximately $(32.5, 45.0)$ pixels.

We characterise the impact of this discrepancy through four diagnostic experiments on validation-split cutouts.

Normalisation statistics ($n = 1000$). The old annulus yields a median offset of $+0.000345$ nmgy relative to the corrected annulus, corresponding to 0.15 normalised units (1.5 per cent of the clip range). The MAD is unchanged (KS test $p = 0.648$). The offset shows no correlation with PSF FWHM ($r = -0.025$, $p = 0.43$) or depth ($r = 0.026$, $p = 0.42$).

Mismatched scoring ($n = 500$ positives $+500$ negatives). Scoring validation cutouts with the corrected annulus (mismatched to the training annulus) yields a recall drop of 3.6 pp at $p > 0.3$ ($z = 1.27$, $p = 0.10$, not significant). This is consistent with the expected sensitivity of any neural network to changes in its input distribution.

Split balance. Two-sample KS tests confirm that PSF FWHM ($p = 0.174$) and depth ($p = 0.123$) distributions are balanced between training and validation splits, ensuring the annulus discrepancy affects both sets equally.

Spatial integrity. Recomputed HEALPix assignments confirm zero Tier-A spatial overlap between training and validation sets (274 and 112 unique pixels, respectively).

We conclude that the annulus discrepancy is cosmetic for model performance. It does not bias the relative comparison between real and injected lenses (both are processed through the same normalisation), and it does not introduce condition-dependent distortions across the survey footprint. We retain the training-consistent annulus for all analyses in this paper.

This paper has been typeset from a TeX/LaTeX file prepared by the author.