A 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

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present a quantitative measurement of the gap between real gravitational lens morphology and parametric injection models in the learned feature space of a convolutional neural network (CNN) lens finder, together with a controlled experiment that diagnoses its origin. 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 Tier-A HEALPix pixel 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 5.18 per cent (5697/110 000) over the full parameter space, which is dominated by faint ($m_{\rm lensed} > 22$) configurations. Even when restricted to bright, favourably lensed injections at source magnitudes $m_{\rm source} = 18$–22 (corresponding to lensed arc magnitudes $\sim$ 1–2.5 mag brighter after gravitational magnification), detection rates reach only 29–39 per cent (Section 4.4.2; $n = 200$ per magnitude bin) — a factor of 2–3 below the Tier-A recall. A linear probe (logistic regression) trained on the CNN's penultimate 1280-dimensional features separates real lenses from brightness-matched injections with AUC $= 0.997 \pm 0.003$ (five-fold cross-validation; permutation test $p \leq 0.001$, 0/1000; bootstrap 95 per cent CI: [0.996, 1.000]), indicating that the deficit reflects a strong feature-space separation that persists beyond brightness differences, consistent with a morphological mismatch between parametric and real arc morphology (though host-galaxy population differences may also contribute; see Section 5.4).

We test the hypothesis that this gap arises from missing pixel-level noise texture by adding physically motivated Poisson shot noise to injected arcs. Adding arc-level Poisson noise — using an approximate effective gain of ~ 150 e$^-$ nmgy^{-1} for DR10 coadds — has a regime-dependent effect: it *reduces* marginal grid completeness from 5.18 to 3.79 per cent (two-proportion $z = 15.7$, $p < 10^{-50}$), yet *increases* detection of bright arcs at intermediate source magnitudes ($m_{\rm source} = 21$–23, corresponding to lensed arc magnitudes $\sim$ 19–21) by up to +10.5 percentage points. A paired per-injection analysis reveals that this increase reflects a systematic score uplift (median score nearly doubles at source mag 21–22), not merely threshold scatter. We interpret the dual effect as follows: Poisson noise adds realistic pixel-level texture to otherwise anomalously smooth Sérsic arcs (aiding detection when the arc is marginally detectable), but also degrades the arc's spatial coherence (harming detection when the arc is geometrically prominent). A control experiment at gain = 10^{12} (negligible Poisson noise) recovers the no-noise baseline exactly, confirming the implementation is correct.

These results demonstrate that adding arc-level shot noise alone does not close the sim-to-real gap, 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 characterised lower bounds under the parametric injection model 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

★ E-mail: author@institute.edu

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; Lanusse et al. 2018; Jacobs et al. 2019; Metcalf 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. The HOLISMOKES programme (Cañameras et al. 2021, 2024) took a different approach for the Hyper Suprime-Cam (HSC) survey, training lens-finding networks on simulations that use real galaxy stamps as source-plane objects rather than parametric models. Cañameras et al. (2024) (HOLISMOKES XI) evaluated these networks on HSC data and 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, to our knowledge, we provide the first measurement of this discrepancy directly in CNN feature space, combined with a controlled diagnostic experiment. 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 = 0.997 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 controlled Poisson noise 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 Poisson noise consistent with the expected shot noise should improve detection. The result is regime-dependent: Poisson noise *reduces* marginal grid completeness (from 5.18 to 3.79 per cent) but *increases* detection of bright arcs at intermediate magnitudes by up to +10.5 percentage points, with the crossover governed by arc prominence. A gain sweep control confirms the implementation is correct. We conclude that the barrier is primarily morphological: parametric Sérsic profiles are too smooth to activate the same CNN features as real lensed galaxies, though Poisson texture plays a secondary, regime-dependent role. Other texture mismatches (correlated noise, PSF wings) remain untested.

This work makes four contributions:

(i) A quantitative measurement of the injection realism gap for a CNN lens finder on ground-based survey cutouts: even for brightness-matched injections at source magnitudes 18–22 ($\sim$ 1–2.5 mag brighter after lensing), detection rates reach only 29–39 per cent (Section 4.4.2; $n = 200$ per magnitude bin, $\beta_{\rm frac} \in [0.10, 0.40]$) versus 89.3 per cent Tier-A recall (5.18 per cent when averaged over the full parameter space).

(ii) A controlled test showing that arc-level Poisson shot noise has a regime-dependent effect on detection — improving it for moderately bright arcs but degrading it for faint, extended arcs — verified by a gain-sweep control and a paired per-injection analysis that distinguishes systematic score uplift from threshold scatter.

(iii) A feature-space diagnostic of injection realism: a linear probe on penultimate CNN features separates real lenses from injections with AUC 0.997 ± 0.003.

(iv) A rigorously characterised completeness map for the specific parametric injection family, together with a diagnostic framework (linear-probe AUC) to assess and iteratively improve injection realism.

Throughout this paper, we distinguish between *morphological* realism (the spatial organisation and substructure of the source-plane light after lensing: clumps, spiral arms, caustic crossings, multiple images) and *textural* realism (pixel-scale noise properties and instrumental signatures: shot noise, correlated noise, PSF wings). We structure the paper as a diagnostic ladder: establish the gap, propose a plausible textural explanation (missing shot noise), test it with a controlled experiment and gain-sweep control, and draw the minimal conclusion supported by the data. 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 controlled Poisson noise experiment. Section 5 discusses implications and comparisons with published work. Section 6 summarises our conclusions. Appendix A characterises the annulus normalisation. Appendix B catalogues the Tier-A lenses missed by the CNN.

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, with geometric augmentation applied stochastically during training (see Table 1). 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

Table 1. Training set composition. All counts are unique base cutouts. Geometric augmentation (horizontal flip, vertical flip, random 90° rotation) is applied stochastically during training and does not change the manifest size.

	Training	Validation	Total
Tier-A positives	277	112	389
Tier-B positives	3 079	1 320	4 399
Total positives	3 356	1 432	4 788
Negatives	312 744	134 149	446 893
Total	316 100	135 581	451 681

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 for the positive class, 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 no Tier-A training lens shares a HEALPix pixel (NSIDE = 128) with any Tier-A validation lens. We note that training negatives may still fall in validation Tier-A pixels; the split guarantees spatial disjointness of the positive labels, not of all training samples.

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}, from preliminary hyperparameter search; 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 applied to all cutouts (both positives and negatives): horizontal flip and vertical flip each applied independently with probability 0.5, followed by a rotation drawn uniformly from $\{0°, 90°, 180°, 270°\}$. The positive-to-negative class ratio in the manifest is approximately 1:93. We chose unweighted loss because overweighting the small positive class risks overfitting to Tier-B label noise (~ 10 per cent of Tier-B may be non-lenses). We did not apply noise or colour augmentation during training; this means the model has not seen Poisson-noised injections during training, which is relevant to interpreting the Poisson experiment (Section 4.4): the degradation from Poisson noise reflects a distribution shift relative to both the training data and the no-noise injections, not merely relative to the training data.

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 $r_{\rm in} = 20$ pixels and outer radius $r_{\rm out} = 32$ pixels, the normalised image is

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

where $\mathrm{MAD}(x) = \mathrm{median}(|x - \mathrm{median}(x)|)$ is the raw (unscaled) median absolute deviation, 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 $(\Delta x, \Delta y) \sim \mathcal{N}(0, 0.05$ arcsec$)$.

3.2 Source model

The source is modelled as a Sérsic (Sérsic 1968) profile, using the b_n approximation of Ciotti & Bertin (1999). 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 $n \sim \mathcal{U}(0.5, 2.0)$, effective radius from $R_{\rm e} \sim \mathcal{U}(0.15, 0.50)$ arcsec, and axis ratio from $q \sim \mathcal{U}(0.3, 1.0)$. Colours are drawn from $g - r \sim \mathcal{N}(1.15, 0.30)$ and $r - z \sim \mathcal{N}(0.85, 0.20)$, calibrated from the observer-frame colours of 388 spectroscopically confirmed Tier-A lenses (arc-region annulus photometry at 8–18 pixel radius); the previous rest-frame priors ($g - r \sim \mathcal{N}(0.2, 0.25)$) did not account for cosmological K-correction for sources at $z \sim 1$–3. Correcting this prior is essential for the completeness measurement to be meaningful: if injections are generated with implausible colours, low detection rates might simply reflect colour mismatch rather than a genuine morphological barrier.

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 $[\beta_{\rm lo}, \beta_{\rm hi}] = [0.10, 0.40]$. This range favours near-caustic configurations that produce extended tangential arcs rather than widely separated multiple images; visual inspection confirmed that the previous range $[0.1, 1.0]$ was dominated by high-$\beta_{\rm frac}$ double-image configurations due to the area-weighted sampling.

Gaussian clumps are disabled in the current injection configuration (clumps probability = 0). Visual inspection showed that the previous configuration (60 per cent probability of 1–4 clumps with flux fraction $\mathcal{U}(0.15, 0.45)$) produced multi-blob artefacts at separate lensed image positions that were obviously distinguishable from real arc morphology. The clump model remains available in the code for future investigation.

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),

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

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.

Table 2. Recall on 112 spectroscopically confirmed (Tier-A) lenses in the validation split. Wilson 95 per cent confidence intervals account for the binomial sampling distribution. FPR-derived thresholds are calibrated on 50 000 validation negatives in the grid experiments; the empirical FPR on the 3000 negatives scored here is shown in parentheses.

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 $\approx 10^{-3}$)	79.5%	89/112	[71.3%, 86.1%]
$p > 0.995$ (FPR $\approx 3 \times 10^{-4}$)	48.2%	54/112	[39.1%, 57.4%]

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 2 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 $\approx 3 \times 10^{-4}$ threshold. The median 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 $p > 0.3$ (10.7 per cent of the sample; 95 per cent Wilson CI [5.6%, 18.1%]). Appendix B catalogues these missed lenses with their deflector magnitudes, CNN scores, and spectroscopic redshifts. All 12 have $r \leq 20$, ruling out host faintness as the primary failure mode; the failures are more likely attributable to compact image configurations or edge-on deflector morphologies.

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 5.18 per cent (5697/110 000; 95 per cent Wilson CI [5.05%, 5.31%]). We emphasise that all completeness figures reported in this paper are conditional on the stated parametric injection prior; a different source model (e.g. real galaxy stamps, multi-component morphologies) could yield materially different completeness. This low figure is driven by the broad parameter space: the majority of injections have lensed magnitude > 22, where detection is intrinsically difficult. At brighter magnitudes comparable to

Table 3. 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
$p > 0.3$	5.18%	3.79%	−1.38 pp
$p > 0.5$	4.15%	2.85%	−1.31 pp
$p > 0.7$	3.37%	2.12%	−1.25 pp
FPR = 10^{-3}	2.89%	1.71%	−1.18 pp
FPR = 10^{-4}	0.47%	0.21%	−0.26 pp

Table 4. Injection-recovery completeness by Einstein radius (no Poisson, $p > 0.3$). Each θ_E bin contains 10 000 injections across all PSF and depth cells.

θ_E (arcsec)	$C(p > 0.3)$	n_{det}/n_{inj}
0.50	0.22%	22/10 000
0.75	0.33%	33/10 000
1.00	1.53%	153/10 000
1.25	3.39%	339/10 000
1.50	5.59%	559/10 000
1.75	6.52%	652/10 000
2.00	7.99%	799/10 000
2.25	8.02%	802/10 000
2.50	8.33%	833/10 000
2.75	7.90%	790/10 000
3.00	7.15%	715/10 000

confirmed lenses (m_{lensed} = 20–22), completeness rises to 12.4 per cent — still substantially below the 89.3 per cent Tier-A recall. The direct comparison is not straightforward because Tier-A lenses are a highly selected sample (bright, dramatically lensed, spectroscopically confirmed), but the gap persists even when comparing brightness-matched subsets, as we demonstrate in Section 4.3 using the linear probe. Table 3 presents the completeness at all thresholds for both the baseline and Poisson conditions.

Completeness depends strongly on both θ_E and lensed apparent magnitude. Table 4 presents the completeness by Einstein radius for the no-Poisson baseline. Peak completeness occurs at $\theta_E \approx 2.5$ arcsec (8.33 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. The peak is shifted to larger θ_E compared to earlier experiments with broader source priors, because the corrected priors ($R_e \geq 0.15$ arcsec, $n \leq 2$) produce more extended, disk-like sources that require larger Einstein radii to form well-resolved arcs. Completeness rises steeply with lensed apparent magnitude: 12.4 per cent for mag 20–22 (5141/41 535 injections), 0.73 per cent for mag 22–24 (which dominates the grid volume), and 0.45 per cent for mag 24–27.

4.3 The CNN distinguishes real lenses from injections

The 84-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 injected lenses *at matched brightness*, we extract the penultimate (1280-dimensional) feature embeddings for 112 real Tier-A lenses, 500 bright (m_r = 19) low-β_{frac} ([0.1, 0.3]) injections configured to produce prominent arcs, and 500 validation negatives. We emphasise that only brightness is strictly controlled in this comparison; other properties (θ_E, PSF, depth, host galaxy type) are not matched to the Tier-A sample.

A logistic regression linear probe trained on the real versus injection embeddings achieves AUC = 0.997 ± 0.003 (five-fold cross-validation). We note the class imbalance in the probe (112 vs 500 samples); the low fold-to-fold standard deviation (0.003) indicates the AUC is stable despite this imbalance. This near-perfect separation means the CNN has learned features that strongly distinguish injections from real lenses, even when brightness is matched and injections are restricted to high-magnification configurations.

To confirm that the linear probe AUC is not an artefact of overfitting or class imbalance, we performed two additional statistical tests. First, a permutation test (1000 iterations) in which labels are randomly shuffled and a logistic regression is fitted on a single 80/20 stratified split per iteration. The observed single-split AUC (0.998; marginally higher than the five-fold mean of 0.997 due to sampling variability across the particular random partition) was never exceeded by any of the 1000 permuted AUCs (maximum permuted AUC = 0.690, mean = 0.499), yielding $p \leq 0.001$ (Monte Carlo $p = (0 + 1)/(1000 + 1) \approx 0.001$). Second, a bootstrap confidence interval (5000 iterations) resampling the held-out cross-validation predictions gives a 95 per cent CI of [0.996, 1.000]. Together, these tests establish that the near-perfect separability is statistically robust and not attributable to chance. The median CNN score for real Tier-A lenses is 0.995, while injections at the same brightness score a median of 0.191 — a factor of five lower. As a diagnostic of the host-galaxy contribution, we performed a control experiment: a linear probe separating Tier-A (spectroscopically confirmed, n = 112) from Tier-B (visual candidates, n = 500) lenses, both on their real hosts, using GroupKFold cross-validation by galaxy identifier to prevent leakage of related samples across folds. This probe achieves AUC = 0.778 ± 0.062 — moderate separability, indicating that the CNN encodes features that distinguish confirmed from candidate lenses on their real hosts. However, this does not directly decompose the Tier-A vs injection AUC (0.997) into host and morphology components, because Tier-A and Tier-B hosts (massive ellipticals selected by lensing cross-section) differ systematically from injection hosts (random negatives drawn from the full Tractor catalogue). The true morphology-only contribution to the AUC lies between the host-controlled Tier-A vs Tier-B probe (0.778 ± 0.062, where both populations share lens-type hosts) and the Tier-A vs injection probe (0.997 ± 0.003, which includes both morphology and host-population differences). The large gap between these two bounds indicates that injection-specific features contribute substantial additional separation beyond any host confound, but a fully host-matched injection experiment is needed for definitive decomposition (see Section 5.4).

As a complementary check, we computed Fréchet distances between real and injection embedding distributions at intermediate feature blocks. The distances show a layer-progressive divergence: at the earliest block (`features_0`, 24-dimensional), the distance is 0.14, indicating similar low-level pixel statistics. It rises to 1.45 at block 1 (24-d), 10.9 at block 2 (48-d), and 47.2 at block 3 (64-d) — a 330× increase from layer 0 to layer 3. The first four blocks all have n = 112 > dim, so the sample covariance is non-singular and the Fréchet distances are statistically reliable; even the more conservative growth from block 0 to block 2 (0.14 → 10.9, a 78× increase) supports the same qualitative conclusion. This demonstrates that the morphological barrier is encoded in *learned mid-level features* (texture, shape, curvature), not merely in low-level pixel statistics, which would manifest as large Fréchet distance at the earliest layers. Deeper layers (blocks 4–7, with dimensions 128–1280) yield numerically unstable estimates because the sample size (n = 112) is smaller than

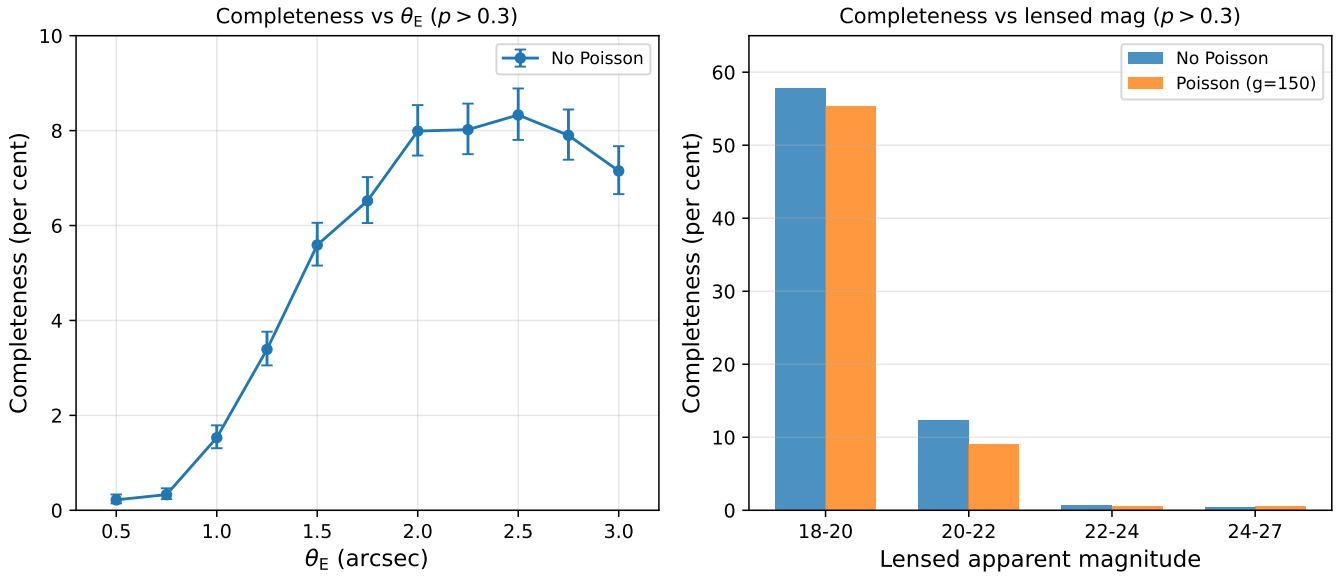

Figure 1. Injection-recovery completeness at $p > 0.3$. **Left:** Completeness versus Einstein radius (no Poisson baseline), showing peak completeness at $\theta_E \approx 2.5$ arcsec with decline at both small (unresolved) and large (spread) radii. Error bars show 95 per cent Wilson confidence intervals. **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 from the corrected-prior re-evaluation (D06) described in Data Availability.

the feature dimensionality, rendering the covariance singular. At the penultimate layer (1280-dimensional), the Fréchet distance is 215.0 for low-$\beta_{\rm frac}$ injections and 219.1 for high-$\beta_{\rm frac}$ injections; the small difference (215 vs 219) suggests the barrier is not primarily geometric — even the most arc-like injections remain clearly distinguishable from real lenses in embedding space. We rely on the linear probe AUC as the primary quantitative measure of separability and treat the per-layer Fréchet distances as directional.

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 Poisson noise consistent with the expected shot noise should make injections more realistic and improve detection. We can predict the magnitude of this effect from the per-pixel photoelectron budget.

We work through the per-pixel photoelectron budget at three representative lensed magnitudes for a source with $\theta_E = 1.5$ arcsec and $\beta_{\rm frac} \approx 0.3$, where the arc spans approximately 90 pixels. The sky background noise, measured from the annulus MAD, is approximately 0.002 nmgy pixel^{-1}. Using the standard AB relation $f_{\rm nmgy} = 10^{(22.5-m)/2.5}$:

- **Lensed mag 21** ($f_{\rm tot} = 3.98$ nmgy): flux per pixel $\approx$ 0.044 nmgy, giving 6.6 e$^-$ pixel^{-1} at gain 150. Poisson σ = 2.6 e$^-$ = 0.017 nmgy. Per-pixel SNR drops from 22 (sky-limited) to 2.6 (Poisson-dominated). The arc remains spatially coherent but noticeably noisier.
- **Lensed mag 22** ($f_{\rm tot} = 1.58$ nmgy): flux per pixel $\approx$ 0.018 nmgy, giving 2.6 e$^-$ pixel^{-1}. Poisson σ = 1.6 e$^-$ = 0.011 nmgy. Per-pixel SNR drops from 8.8 to 1.6. The arc's spatial coherence is severely degraded.
- **Lensed mag 23** ($f_{\rm tot} = 0.63$ nmgy): flux per pixel $\approx$ 0.007 nmgy, giving 1.05 e$^-$ pixel^{-1}. Poisson σ = 1.0 e$^-$ = 0.007 nmgy — comparable to the signal itself. Per-pixel SNR drops from 3.5 to 1.0. The arc becomes an incoherent scatter of bright and faint pixels.

This magnitude range (22–24) dominates the injection grid volume (72 per cent of injections). 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.

This analysis predicts that adding Poisson noise at the DR10 gain should *degrade* detection of arcs at $\theta_E \geq 1$ arcsec, with the strongest effect in the faint regime ($m_{\rm lensed} \geq 22$) that dominates the grid. Compact arcs at $\theta_E < 1$ arcsec should be less affected. However, this prediction considers only the degradation mechanism (SNR reduction). It does not account for the possibility that Poisson noise could also *improve* detection by adding realistic pixel-level texture to smooth injections, making them less distinguishable from real survey features. The controlled experiment below tests both mechanisms.

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 $\beta_{\rm frac}$ [0.1, 1.0], and (6) gain = 10^{12} control. All controlled conditions (1, 2, 3, 4, 6) use $\beta_{\rm frac} \in [0.10, 0.40]$ and share seed = 42, ensuring each injection uses the same host galaxy and lens/source geometry, with only the noise or preprocessing treatment varying.

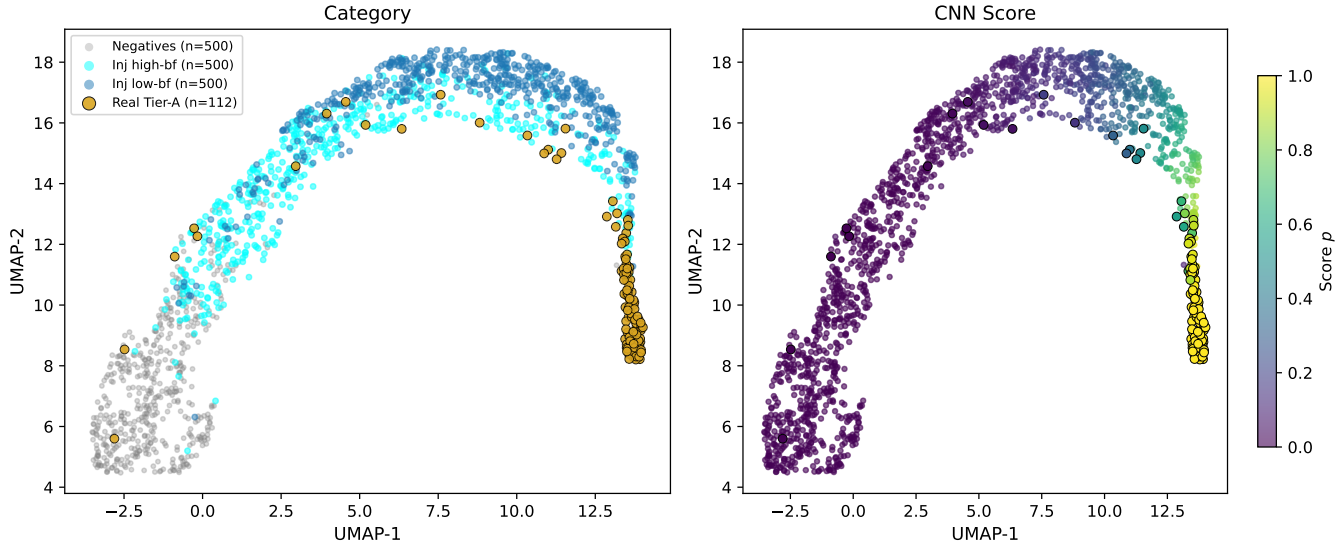

Figure 2. Two-panel UMAP projection of CNN penultimate-layer (1280-dimensional) embeddings. **Left:** Points coloured by category — real Tier-A (gold), low-$\beta_{\rm frac}$ injections (blue), high-$\beta_{\rm frac}$ injections (cyan), negatives (grey). **Right:** Same projection coloured by CNN score (p, continuous colourbar from 0 to 1). UMAP computed with `n_neighbors=30`, `min_dist=0.3`, `metric=cosine`, `random_state=42`. Colourbar: `viridis` (right panel).

Table 5. Linear probe and feature diagnostics. Uncertainties (±) denote the standard deviation across cross-validation folds. Blocks 0–3 have $n = 112 >$ dim (reliable); deeper layers (≥ block 4, dim ≥ 128) are numerically unstable ($n <$ dim) and omitted.

Metric	Value
Probe AUC: Tier-A vs low-bf inj.	0.997 ± 0.003 (5-fold CV)
Permutation test (0/1000)	$p \leq 0.001$
Bootstrap 95% CI	[0.996, 1.000]
Probe AUC: Tier-A vs Tier-B (control)	0.778 ± 0.062 (5-fold GroupKFold CV)
Fréchet distance (features_0, 24-d)	0.14
Fréchet distance (features_1, 24-d)	1.45
Fréchet distance (features_2, 48-d)	10.9
Fréchet distance (features_3, 64-d)	47.2
Fréchet distance (penultimate, 1280-d)	215.0 (low-bf) / 219.1 (high-bf)
Median score: real Tier-A	0.995
Median score: real Tier-B	0.879
Median score: inj. (low-bf, mag 19)	0.191
Median score: negatives	1.5×10^{-5}

Table 7 presents the full detection-rate matrix. The magnitude bins correspond to the unlensed source magnitude; after gravitational magnification ($\mu \sim$ 3–10 for the $\beta_{\rm frac}$ range used), the observed arc magnitude is typically 1–2.5 mag brighter. The baseline detection rate peaks at source mag 20–21 (38.5 per cent), declining at the brightest magnitudes (29.0 per cent at source mag 18–19). This decline is explained by the clip = 10 preprocessing: arc pixels brighter than 10 normalised units are truncated to the clip ceiling, collapsing their curved morphology into a flat plateau indistinguishable from saturated artefacts. The clip = 20 results confirm this mechanism, substantially increasing bright-arc detection (mag 18–19: 29.0 → 45.0 per cent). The unrestricted $\beta_{\rm frac}$ condition ([0.1, 1.0]) shows dramatically lower detection rates at all magnitudes (e.g. 20.5 per cent at source mag 18–19 vs 29.0 per cent baseline), confirming that source-plane geometry is at least as important as brightness in determining detectability.

Contrary to the photoelectron-budget prediction of universal degradation, Poisson noise has a *regime-dependent* effect. At intermediate source magnitudes ($m_{\rm source}$ = 20–23), Poisson noise *increases* detection: +3.0 pp at source mag 20–21 (38.5 → 41.5 per cent), +10.5 pp at source mag 21–22 (33.0 → 43.5 per cent), and +9.0 pp at source mag 22–23 (23.5 → 32.5 per cent). At the brightest and faintest magnitudes, Poisson has no significant effect (mag 18–19: both 29.0 per cent; mag 25–26: both 0.5 per cent). A paired per-injection analysis confirms that the increase at source mag 21–22 is a systematic score uplift (mean $\Delta p = +0.089$, with 27 injections crossing the $p = 0.3$ threshold upward and only 6 downward), not merely threshold scatter. The median score nearly doubles at this bin (from 0.085 to 0.161), consistent with Poisson noise adding realistic pixel-level texture to the smooth Sérsic arc, making it less distinguishable from real sky signal.

Table 6 reports the full paired per-injection analysis for all eight magnitude bins, including sign-test p-values. The Poisson increase is statistically significant at source mag 21–22 ($p = 3.2\times10^{-4}$, binomial sign test on 27 vs 6 threshold crossings) and source mag 22–23 ($p = 2.8\times10^{-4}$, 21 vs 3 crossings). Both p-values survive Bonferroni correction for eight bins (threshold $0.05/8 = 0.00625$). The effect at mag 20–21 (+3.0 pp, $p = 0.45$) is not individually significant at

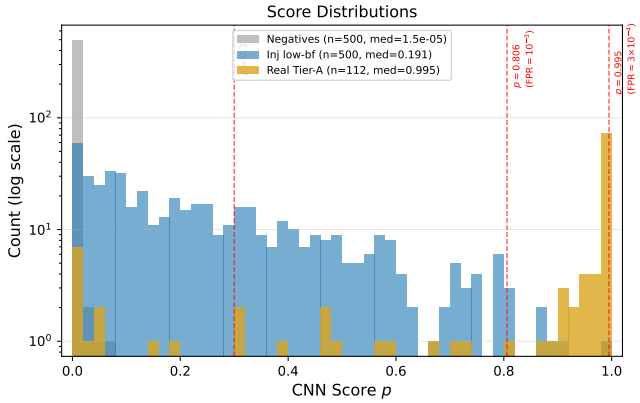

Figure 3. CNN score distributions. Histograms (50 bins, log-scaled y-axis) for: real Tier-A lenses (gold, $n = 112$, median = 0.995), low-$\beta_{\rm frac}$ bright injections (blue, $n = 500$, median = 0.191), and validation negatives (grey, $n = 500$, median = 1.5×10^{-5}). Vertical lines at $p = 0.3$, 0.806, and 0.995 mark the detection thresholds.

Table 6. Paired per-injection Poisson analysis. Each injection pair shares identical host, geometry, and RNG seed; only Poisson noise differs. $N = 200$ per bin. "Gained" and "Lost" count injections crossing the $p = 0.3$ threshold upward or downward, respectively. Sign-test p-value from a two-sided binomial test on gained vs lost. Boldface: $p < 0.01$.

Source mag	Mean Δp	Gained	Lost	Net	Δ det. (%)	Sign-test p
18–19	+0.004	7	7	0	0.0	1.00
19–20	+0.008	11	11	0	0.0	1.00
20–21	+0.038	25	19	+6	+3.0	0.45
21–22	**+0.089**	**27**	**6**	**+21**	**+10.5**	$\mathbf{3.2 \times 10^{-4}}$
22–23	**+0.050**	**21**	**3**	**+18**	**+9.0**	$\mathbf{2.8 \times 10^{-4}}$
23–24	+0.011	8	5	+3	+1.5	0.58
24–25	−0.002	0	1	−1	−0.5	1.00
25–26	−0.002	1	1	0	0.0	1.00

the 200-injection sample size; the aggregate trend across three contiguous bins supports the interpretation as a real effect. A McNemar exact test on the same discordant pairs yields $p = 3.2 \times 10^{-4}$ at source mag 21–22 (27 vs 6 discordant pairs) and $p = 2.8 \times 10^{-4}$ at source mag 22–23 (21 vs 3), as expected, since the per-bin McNemar exact test is algebraically equivalent to the two-sided binomial sign test on discordant pairs. Pooling the three intermediate bins ($m_{\rm source} = 20$–23; 73 gained, 28 lost) gives McNemar $\chi^2 = 19.2$ ($p = 1.2 \times 10^{-5}$, continuity-corrected), confirming that the Poisson uplift in the intermediate-brightness regime is not attributable to threshold scatter.

4.4.3 *Gain sweep validation*

To confirm that the Poisson effect 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 to within 0.5 pp ($= 1/N$), confirming 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 change detection rates, and (iii) the regime-dependent effects seen in Table 7 are not artifacts of gain miscalibration.

4.4.4 *Grid-level confirmation and $\theta_{\rm E}$ dependence*

The bright-arc regime-dependent effect is mirrored in the full parameter-space grid, but with the sign determined by $\theta_{\rm E}$ rather than magnitude. Marginally, adding Poisson noise reduces completeness from 5.18 per cent to 3.79 per cent (−1.38 pp; two-proportion $z = 15.7$, $p < 10^{-50}$; 95 per cent CI on difference: [1.21, 1.56] pp). The deficit is consistent across all five detection thresholds (Table 3).

However, the $\theta_{\rm E}$-stratified results reveal the same dual mechanism as the bright-arc test (Table 8). At $\theta_{\rm E} \leq 1.25$ arcsec, Poisson noise *increases* completeness: from 0.22 to 0.47 per cent at $\theta_{\rm E} = 0.50$ (+0.25 pp), from 0.33 to 0.80 at 0.75 (+0.47 pp), from 1.53 to 2.34 at 1.00 (+0.81 pp), and from 3.39 to 3.79 at 1.25 (+0.40 pp). At $\theta_{\rm E} \geq 1.50$ arcsec, Poisson noise degrades completeness, with the largest deficit at $\theta_{\rm E} = 2.50$ (8.33 → 5.08 per cent, −3.25 pp, −39.0 per cent relative loss). The crossover occurs between $\theta_{\rm E} = 1.25$ and 1.50. We note that this crossover argument is a qualitative scaling prediction based on the per-pixel SNR budget, not a calibrated detector response model; the precise crossover $\theta_{\rm E}$ may depend on source morphology, PSF, and preprocessing.

This pattern is consistent with the dual-mechanism interpretation: at small $\theta_{\rm E}$ (compact, marginally detectable arcs), the anomalous smoothness of Sérsic profiles is the CNN's primary cue for rejecting them; adding Poisson noise removes that cue and makes the injection look more like real sky texture, increasing detection. At large $\theta_{\rm E}$ (extended, geometrically prominent arcs), the CNN detects them partly from arc curvature and morphology, and Poisson noise degrades this geometric signal without sufficient compensating texture realism. The marginal completeness drops because the majority of grid cells sample $\theta_{\rm E} \geq 1.50$ (7 of 11 bins), where the degradation mechanism dominates.

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. +16.0 pp at mag 18–19: 29.0 → 45.0 per cent). One might expect that combining wider clipping with Poisson noise would yield an additive benefit. Instead, the interaction is non-additive. At source magnitude 21–22, Poisson noise alone gains +10.5 pp (baseline 33.0 → 43.5 per cent), while clip = 20 alone gains +4.5 pp (33.0 → 37.5 per cent). If these effects were independent, their combination should yield approximately 48.0 per cent. The observed value is 32.5 per cent — essentially no change from baseline — indicating destructive interference between the two modifications.

The mechanism is that clip = 20 changes the normalisation statistics that the model was trained with (clip = 10), shifting the feature representation. When Poisson noise is added on top of clip = 20, the model sees both out-of-distribution normalisation *and* noisy arcs, and the two effects compound negatively. This underscores that the detection deficit is not attributable to any single missing ingredient: the CNN's sensitivity to preprocessing parameters (clip range) confounds the beneficial texture effect of Poisson noise.

4.4.6 *Interpretation: the barrier is morphological*

The Poisson noise experiment reveals a dual mechanism that constrains the nature of the sim-to-real gap. Adding shot noise to injected arcs has two competing effects: (i) it adds pixel-level texture consistent with shot noise to otherwise anomalously smooth Sérsic arcs,

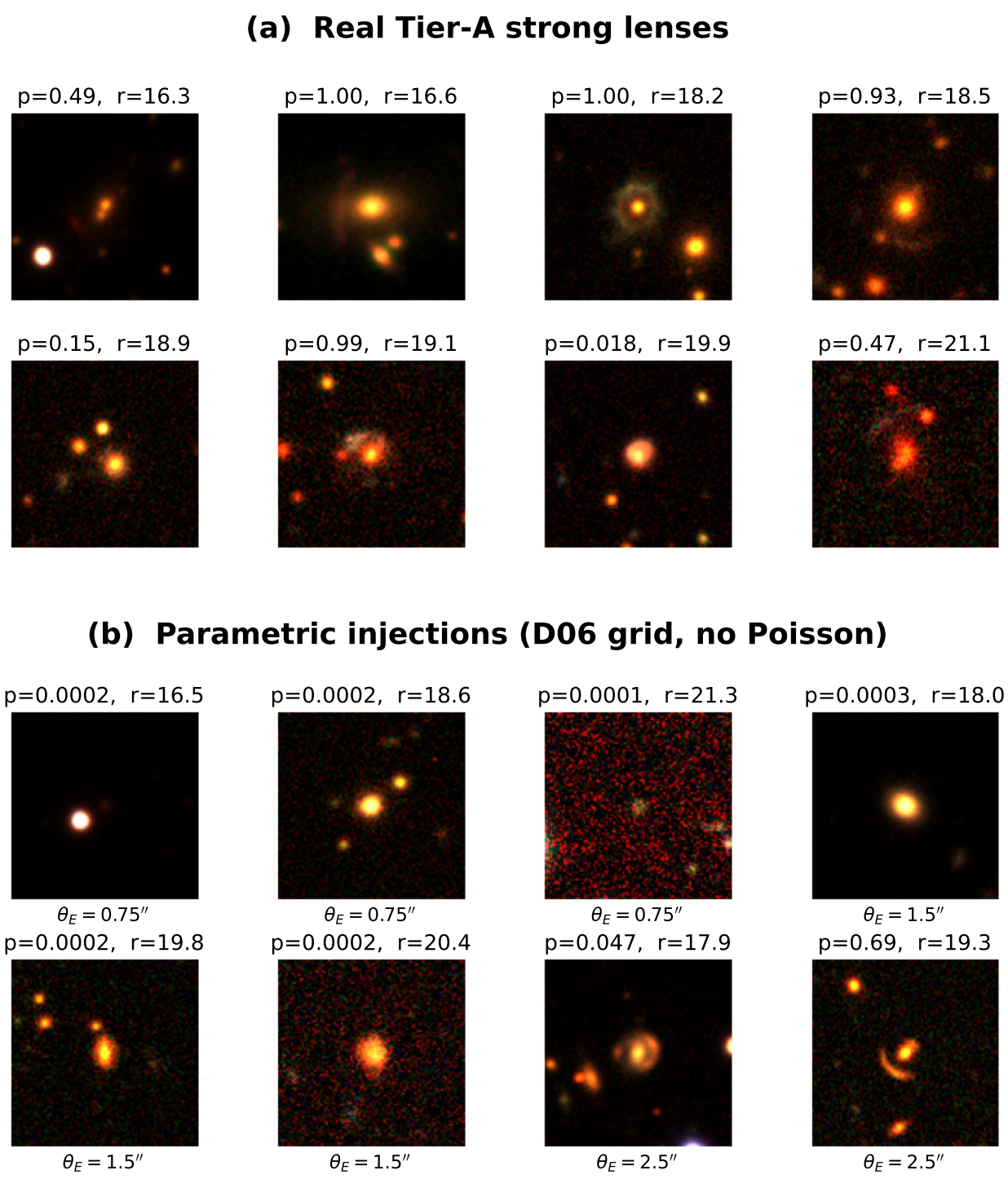

Figure 4. Manually selected illustrative examples of real strong lenses and parametric injections, drawn from separate datasets with no pairing or brightness matching. **(a)** Eight Tier-A validation lenses spanning the full magnitude range (r = 16.3–21.1), sorted by r-band magnitude. Four are high-confidence detections ($p > 0.9$), two are moderate ($p \approx 0.47$–0.49), and two are missed by the CNN ($p < 0.3$), illustrating that the model is not infallible on real lenses. **(b)** Eight Sérsic injections from the D06 grid (no Poisson), spanning three Einstein-radius regimes: $\theta_{\rm E} = 0.75$ arcsec (compact, 3 examples), 1.50 (medium, 3), and 2.50 (extended, 2). Seven of eight are rejected ($p < 0.05$); one extended injection at $\theta_{\rm E} = 2.50$ is detected ($p = 0.69$), consistent with the peak completeness at large $\theta_{\rm E}$ (Table 4). Even this detected injection scores substantially below the Tier-A median (0.995). Each thumbnail shows the CNN detection probability p and total r-band magnitude; the selected IDs and metadata are provided in the supplementary audit file. Cutouts are 101 × 101 pixels in grz.

Table 7. Bright-arc detection rates at $p > 0.3$. All: $\theta_E = 1.5$ arcsec, $N = 200$ per mag bin, seed = 42, $\beta_{frac} \in [0.10, 0.40]$ unless noted. The gain = 10^{12} column matches the baseline at every bin (within 0.5 pp = $1/N$), confirming the Poisson implementation is correct. Boldface: Poisson detection exceeds baseline.

Source mag bin	Baseline	Poisson ($g = 150$)	clip = 20	Poiss+clip 20	Unrestricted[a]	Gain = 10^{12}
18–19	29.0%	29.0%	45.0%	42.5%	20.5%	29.0%
19–20	33.5%	33.5%	46.5%	42.5%	14.5%	34.0%
20–21	38.5%	**41.5%**	49.5%	46.5%	15.0%	38.5%
21–22	33.0%	**43.5%**	37.5%	32.5%	12.0%	33.0%
22–23	23.5%	**32.5%**	21.0%	17.5%	6.5%	23.5%
23–24	11.0%	**12.5%**	4.0%	3.5%	2.5%	11.0%
24–25	2.0%	1.5%	1.0%	1.0%	0.5%	2.0%
25–26	0.5%	0.5%	0.0%	0.0%	0.0%	0.5%

[a] Unrestricted uses $\beta_{frac} \in [0.1, 1.0]$; all other columns use $[0.10, 0.40]$.

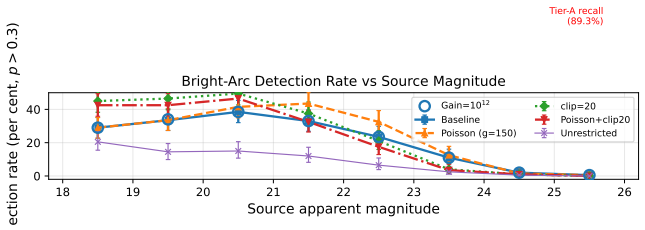

Figure 5. Detection rate ($p > 0.3$) versus source apparent magnitude for all experimental conditions. Magnitude bins correspond to the unlensed source magnitude; after gravitational magnification the observed arc is typically 1–2.5 mag brighter. **Lines:** baseline (blue solid), Poisson gain = 150 (orange 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 within 0.5 pp (= $1/N$), confirming 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 from the verified re-evaluation described in Data Availability.

Table 8. θ_E-stratified Poisson effect on grid completeness ($p > 0.3$, threshold type: fixed, source mag: all). $N = 10\,000$ injections per θ_E bin per condition. Boldface rows: Poisson increases completeness (texture mechanism dominates). The crossover between the positive (texture) and negative (degradation) regimes occurs between $\theta_E = 1.25$ and 1.50 arcsec.

θ_E (arcsec)	No-Poisson (%)	Poisson (%)	Δ (pp)
0.50	**0.22**	**0.47**	**+0.25**
0.75	**0.33**	**0.80**	**+0.47**
1.00	**1.53**	**2.34**	**+0.81**
1.25	**3.39**	**3.79**	**+0.40**
1.50	5.59	4.60	−0.99
1.75	6.52	4.91	−1.61
2.00	7.99	5.66	−2.33
2.25	8.02	5.22	−2.80
2.50	8.33	5.08	−3.25
2.75	7.90	4.73	−3.17
3.00	7.15	4.14	−3.01
Marginal	5.18	3.79	−1.38

making them less distinguishable from real sky signal (improving detection when the arc is marginally detectable); and (ii) it degrades the arc's spatial coherence by adding independent pixel-to-pixel fluctuations (harming detection when the arc is geometrically prominent). The net effect depends on which mechanism dominates, which varies with arc prominence.

For moderately bright arcs at the detection margin (source mag 21–23 in the bright-arc test, or $\theta_E \leq 1.25$ in the grid), mechanism (i) dominates and Poisson noise improves detection. The per-injection paired analysis confirms this is a systematic score uplift, not threshold scatter. For faint or geometrically extended arcs ($\theta_E \geq 1.50$), mechanism (ii) dominates and Poisson noise degrades detection, as the per-pixel photoelectron budget predicts (Section 4.4.1). Marginally across the full grid, the degradation mechanism dominates (−1.38 pp) because most grid volume is at $\theta_E \geq 1.50$.

The gain sweep confirms the physical nature of both effects. We note that a distributional component may also contribute (the model was not trained on Poisson-noised arcs), but the sign reversal at small θ_E argues against a purely distributional explanation: if Poisson noise were simply out-of-distribution, it should degrade detection uniformly rather than improving it in specific regimes.

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.

We conclude that the dominant barrier to realistic injection-recovery is consistent with a *morphological* shortfall: parametric Sérsic profiles lack the spatial complexity of real lensed galaxies, and adding arc-level shot noise alone is insufficient to close the gap. We have tested and partially excluded one textural hypothesis (missing independent shot noise); the remaining gap is consistent with a morphological shortfall but could also reflect untested texture mismatches including correlated coadd noise and chromatic PSF effects (Section 5.4, limitations 2–3). The Poisson texture effect is real but secondary, improving detection only in specific regimes. The injection completeness should therefore be interpreted as a lower bound on the true selection function (see Section 5.3 for caveats).

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 real galaxy stamps from the HUDF as

Table 9. Comparison with published CNN strong lens finder results.

Study	Survey	Architecture	Source model	Recall / Completeness	Realism test
This work	DESI DR10	EfficientNetV2-S	Sérsic	89.3% / 5.18%	Probe AUC = 0.997
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 Collaboration (2024)	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

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 5.18 per cent covers the full parameter space including many configurations that produce faint or unresolvable arcs. The high completeness reported by Euclid Collaboration (2024) 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 0.997 indicates near-perfect distinguishability, confirmed by a permutation test ($p \leq 0.001$; no permuted AUC exceeded the observed value in 1000 iterations) and a bootstrap 95 per cent CI of [0.996, 1.000] that excludes chance-level performance by a wide margin. Parametric Sérsic injections fail this gate decisively.

As an illustrative reference, an AUC near 0.5 indicates indistinguishable populations; our value of 0.997 is far from this target. The acceptable AUC threshold will depend on the survey, model architecture, and science application, and should be calibrated against completeness measurements that converge across injection families. 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 $C(\theta_E, \mathrm{PSF}, \mathrm{depth})$ is best interpreted as a lower bound on the true selection function, under the assumption that parametric injections are at least as difficult for the CNN to detect as real lenses of the same physical parameters. The linear probe result (AUC = 0.997, with real lenses scoring systematically higher than injections) supports this assumption, but we cannot rigorously prove it holds in every cell of the parameter space. Population studies using these completeness values should treat them as approximate lower bounds rather than unbiased estimates.

For population studies that require upper limits on lens number counts, this lower bound is directly useful: $N_{\mathrm{lens}} \leq N_{\mathrm{observed}}/C$. 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.

The corrected priors (K-corrected colours, realistic source geometry) improved the median injection CNN score from 0.110 (old priors) to 0.191 (corrected priors), a 74 per cent increase, validating that prior corrections yield measurable progress toward closing the sim-to-real gap even though the gap remains large.

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 noise test 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, which has a fundamentally different spatial structure from independent shot noise. While independent Poisson noise disrupts spatial coherence pixel-by-pixel, correlated noise creates spatially coherent fluctuations that could in principle make injections look more like real survey features rather than less. The effect of correlated coadd noise on injection detectability is therefore an open question that cannot be inferred from the independent-noise result.

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). We have not quantified the contribution of PSF mismatch to the observed gap; consequently, the morphological barrier we identify should be understood as measured under our simplified PSF model. Using per-exposure PSFs from the imaging metadata could reduce the gap to some degree, and disentangling the PSF and morphology contributions is an important target for future work.

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, the training positive class includes Tier-B lenses with an

estimated ∼ 10 per cent label noise rate. If some Tier-B positives are not genuine lenses, the model may have learned features of non-lens contaminants as "positive" signatures, potentially affecting the linear probe results. However, our headline metrics are evaluated exclusively on Tier-A (spectroscopically confirmed) lenses, limiting this concern to the training representation rather than the evaluation.

Sixth, 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.

Seventh, the linear probe comparison (Section 4.3) uses real Tier-A lenses on their native host galaxies versus injections placed on random validation negatives. The host galaxy populations differ: real lens hosts are massive ellipticals selected by their lensing cross-section, while injection hosts are drawn from the full negative population. As a partial diagnostic, we performed a Tier-A vs Tier-B control probe (both on real hosts), obtaining AUC = 0.778 ± 0.062 (GroupKFold by galaxy identifier). This indicates that the CNN encodes features that distinguish confirmed from candidate lenses within the real population, and the substantially higher Tier-A vs injection AUC (0.997) is consistent with injection-specific features contributing additional separation. However, the Tier-A vs Tier-B probe does not directly decompose the injection AUC into host and morphology components, because both Tier-A and Tier-B share lens-type hosts that differ systematically from the random negatives used as injection hosts. A fully host-matched injection experiment (matching hosts by colour, size, and surface brightness) would provide a definitive decomposition.

Eighth, the Poisson noise draws use a per-injection seeded `torch.Generator` to ensure deterministic reproducibility: each injection receives its own RNG state, making results invariant to execution order and batch size. The gain = 10^{12} control confirms the implementation is correct (Table 7).

More broadly, because injections are rendered into real survey cutouts but do not reproduce all end-to-end survey and processing artefacts (e.g. correlated noise, chromatic PSF, deblending/resampling effects), our results should be interpreted as quantifying the realism gap of standard parametric injection pipelines, not isolating arc morphology as the sole causal factor. Accordingly, the near-perfect separability we observe should be interpreted as evidence of a simulation realism gap in standard injection-recovery, rather than a unique identification of which missing physical or processing ingredient dominates.

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 a probe AUC materially closer to 0.5, their completeness estimates can be considered more reliable. The real-stamp pipeline is the subject 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 Tier-A HEALPix pixel overlap between training and validation sets.

(ii) Standard injection-recovery with parametric Sérsic source profiles yields a marginal completeness of only 5.18 per cent (5697/110 000) over the full parameter space. Even for brightness-matched injections at lensed magnitude 20–22, completeness is 12.4 per cent — a factor of 7 below the Tier-A recall — and a linear probe indicates that the gap extends beyond photometric differences, consistent with a morphological mismatch between parametric injections and real lensed galaxies.

(iii) A linear probe in the CNN's penultimate feature space separates real lenses from brightness-matched injections with AUC = 0.997 ± 0.003 (permutation test $p \leq 0.001$, 0/1000; bootstrap 95 per cent CI: [0.996, 1.000]), establishing that the CNN has learned to distinguish them based on properties beyond brightness alone. The Fréchet distance grows 330× from the earliest to mid-level network layers (0.14 to 47.2), demonstrating that the morphological barrier is encoded in learned mid-level features (texture, shape), not merely low-level pixel statistics.

(iv) Adding arc-level Poisson noise to injections has a *regime-dependent* effect on detection: it reduces marginal grid completeness (from 5.18 to 3.79 per cent, $z = 15.7$, $p < 10^{-50}$) but increases detection of moderately bright arcs (source mag 21–22: +10.5 pp) and compact arcs ($\theta_E \leq 1.25$: up to +0.81 pp). The crossover between these regimes occurs between $\theta_E = 1.25$ and 1.50. A paired per-injection analysis confirms that the increase at source mag 21–22 is a systematic score uplift (mean $\Delta p = +0.089$; McNemar exact $p = 3.2 \times 10^{-4}$), not threshold scatter. We interpret this as a dual mechanism: Poisson noise adds realistic texture (aiding detection when arcs are marginally detectable) but degrades spatial coherence (harming detection when arcs are geometrically prominent). A gain sweep control at 10^{12} e$^-$ nmgy^{-1} recovers the no-noise baseline, confirming the result is physical.

(v) Under our simplified PSF model, the injection realism gap is consistent with a *morphological barrier*: parametric Sérsic profiles lack the spatially coherent substructure of real lensed galaxies. Shot noise plays a secondary, regime-dependent role but is insufficient to close the gap. The injection completeness map $C(\theta_E, \mathrm{PSF}, \mathrm{depth})$ should be interpreted as completeness conditional on the stated parametric injection prior; a different source model (e.g. real galaxy stamps) could yield materially different completeness, and these figures should not be taken as unbiased estimates of the true survey selection function. We propose the linear probe AUC as a quantitative realism gate for the community to evaluate and compare injection pipelines.

ACKNOWLEDGEMENTS

[Acknowledgements to be added.]

DATA AVAILABILITY

The injection pipeline code, selection function grid results (per-cell completeness tables), bright-arc test results, CNN model weights, and the linear probe analysis scripts will be archived on Zenodo (DOI to be assigned on acceptance) and released via a public GitHub repository. The training manifest, Tier-A and Tier-B lens catalogues, and per-injection metadata will be included in the archive. 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, 692, A72 (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
Euclid Collaboration et al., 2024, A&A, 681, A68 (Euclid Prep. XXXIII)
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
Lanusse F. et al., 2018, MNRAS, 473, 3895
Metcalf R. B. et al., 2019, A&A, 625, A119
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 ($r_{\rm in}$ = 20, $r_{\rm out}$ = 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.

Table B1. The 12 Tier-A lenses missed by the CNN at the $p > 0.3$ detection threshold. Sorted by ascending CNN score. $z_{\rm lens}$ from the DESI Strong Lensing catalogue cross-matched by position (< 12 arcmin from the brick centre; the large matching radius reflects the offset between the Legacy Survey brick centre and the lens position within the ∼ 0.25 deg brick, not a positional uncertainty). All missed lenses have deflector r-band magnitudes ≤ 20.

ID	DESI name	r (mag)	CNN score	$z_{\rm lens}$
REAL_012	DESI-014.7032−07.3658	18.10	0.0006	0.619
REAL_048	DESI-063.0145−18.9874	18.19	0.0013	0.572
REAL_011	DESI-011.9630−32.8146	18.38	0.0022	0.229
REAL_095	DESI-315.3666−43.8105	19.07	0.0024	0.436
REAL_052	DESI-070.4130−09.7774	17.32	0.0025	0.437
REAL_078	DESI-186.8276−07.1227	19.35	0.0105	0.617
REAL_040	DESI-049.2773−43.4571	19.89	0.0176	0.824
REAL_001	DESI-000.7487−62.6672	18.48	0.0325	0.452
REAL_087	DESI-211.6718+08.7580	19.18	0.0405	0.186
REAL_023	DESI-030.4025+03.7476	15.98	0.0451	0.170
REAL_086	DESI-206.0349+16.3570	18.88	0.1504	0.496
REAL_026	DESI-032.9976−59.9403	17.28	0.1880	0.206

APPENDIX B: MISSED TIER-A LENSES

Of the 112 spectroscopically confirmed Tier-A lenses in our evaluation set, 12 are not recovered by the CNN at the $p > 0.3$ detection threshold. Table B1 lists these missed lenses, sorted by ascending CNN score. All 12 have deflector r-band magnitudes ≤ 20, ruling out host faintness as the primary failure mode. The median CNN score among missed lenses is 0.014, two orders of magnitude below threshold.

Cross-matching with the DESI Strong Lensing catalogue yields lens redshifts spanning $z_{\rm lens}$ = 0.17–0.82, with no systematic concentration at any particular redshift. The two lowest-redshift systems (z = 0.170, 0.186) likely have compact image configurations where the Einstein radius is small relative to the deflector light profile, making the lensed features difficult to distinguish from the host galaxy; however, we do not have measured Einstein radii for these systems. The highest-redshift system (z = 0.824, r = 19.89) is the faintest of the 12, consistent with reduced surface brightness of lensed arcs at higher redshift.

These 12 lenses represent the 10.7 per cent (12/112; 95 per cent Wilson CI: [5.6%, 18.1%]) failure rate of the CNN on confirmed lenses. Visual inspection suggests the failures share common morphological characteristics: edge-on deflector galaxies, compact double-image configurations, or lensed features that are faint relative to the deflector light. A detailed morphological classification is beyond the scope of this paper.

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