Attention-Based Preprocessing Framework for Improving Rare Transient Classification*

Xinyue Sheng, Matt Nicholl (ARC), Dung Pham, Zichi Zhang (EEECS), Queen's University Belfast, UK

1. Motivations

Data preprocessing is crucial for machine learning models in large sky surveys, particularly for transient detection. The task is complicated by **extreme class imbalance, irregular observation cadences, and inconsistent image quality**. We present efficient, astronomy-specific data augmentation techniques for repairing imaging artefacts, masking unrelated sources, and simulating realistic light curves from unevenly sampled data. Using the **NEEDLE** benchmark (Sheng et al., 2024) for infant SLSNe-I and TDE detection from photometry, we show that these methods yield substantial performance gains.

Spectroscopic confirmed events (appr.)	SN	SLSN-I	TDE
	> 5,000	~270	~106

Final. Completeness & Purity Comparisons

Original

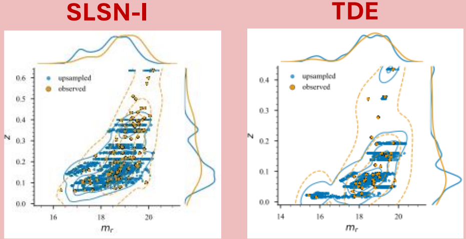
True \ Predicted	SN	SLSN-I	TDE
SN	0.93 0.95	0.048 0.53	0.022 0.31
SLSN-I	0.125 0.005	0.75 0.35	0.125 0.077
TDE	0.474 0.049	0.105 0.12	0.421 0.62

After cross-matching

True \ Predicted	SN	SLSN-I	TDE
SN	0.968 0.96	0.022 0.40	0.011 0.12
SLSN-I	0.25 0.011	0.75 0.60	0 0
TDE	0.263 0.027	0 0	0.737 0.88

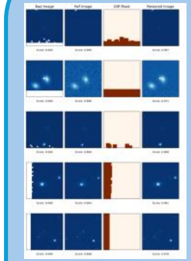
*This is the preliminary results, we are working on more for the paper version!

5. Up-sampled Distribution (SLSN-I & TDE)



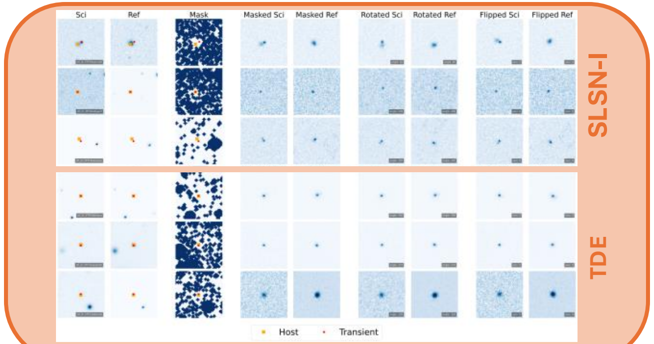
In both plots, we observe that as the redshift increases and the apparent magnitude becomes fainter, the discovery rate declines, which is an expected result of survey observing limits. We can see the same effect is captured in our synthetic data, with the bulk of detections at low redshift.

2. Image Restoration



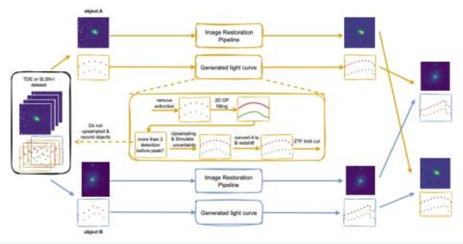
We restore poor-quality Science or Reference images from alerts by identifying good alternatives, padding incomplete images, detecting and replacing bad pixels via **SSIM-based** masks with intensity adjustment, and re-checking quality, enabling reuse of otherwise discarded data for NEEDLE processing.

3. Mask Background Sources & Agumentation



4. Cross-matching for Up-sampling

To solve the small sample problem, we propose up-sampling rare classes (SLSN-I, TDE) by cross-matching host galaxy images with light curves (re-sampled with 2D Gaussian Process) from other objects in the same class, scaled to similar redshifts.


object A
object B
TDE or SLSN-I dataset
Image Restoration Pipeline
Generated light curve
remove extinction
2D GP fitting
more than 2 detection before peak?
Upsampling & Simulate uncertainty
convert A to B redshift
ZTF limit cut
Do not upsampled & record objects

The method generates realistic synthetic samples following observed redshift distributions, boosting dataset size and diversity while retaining astrophysical consistency.

**Sheng et al., 2025 in prep*