Selfish Evolution: Making Discoveries in Extreme Label Noise with the Help of Overfitting Dynamics

Nima Sedaghat[1,2]*, **Tanawan Chatchadanoraset**[1,2], **Colin Orion Chandler**[1,2]

Ashish Mahabal[3], **Maryam Eslami**[2]

[1]Department of Astronomy, University of Washington
[2]Raw Data Speaks Initiative
[3]Department of Astronomy, California Institute of Technology

Abstract

Motivated by the scarcity of proper labels in astrophysical surveys, we introduce Selfish Evolution, a weakly supervised technique that *detects and corrects* corrupted labels in situ. Rather than relying on early stopping, we first train on the noisy dataset and then deliberately overfit to individual samples; the ensuing overfitting dynamics form spatiotemporal "evolution cubes" that are predictive of both label noisiness and its corrected value. A secondary network learns this mapping to produce pixel-level label repairs. The procedure runs in a closed loop—cleaned labels improve the detector, which then reveals fainter events—without assumptions about the model state at intervention. Centered on supernova detection, we demonstrate convergence toward a largely clean training set and recovery of missed astrophysical objects, advancing discovery under extreme label noise.

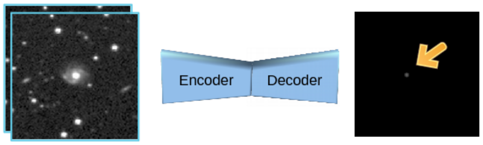

Figure 1: Image-based redefinition of the task of supernova detection. On the left, two images of the same region of the sky are passed to the network, and the output is defined as an image of the same size, containing only the reconstructed desired object Sedaghat and Mahabal [2018]

1 Introduction

Deep learning has become integral to astronomical discovery, particularly to the automatic detection of explosive transients such as supernovae. Traditionally, surveys register and co-add images to form a template and then subtract each new exposure, followed by denoising and detection; the full pipeline hinges on precise alignment, PSF matching, and subtraction quality.

The current state–of–the–art, TransiNet Sedaghat and Mahabal [2018], synthesizes "difference" images by painting the expected transient onto a blank canvas (fig. 1), effectively collapsing the subtraction pipeline into an image–generation task closely related to segmentation, where unwanted pixels are driven to zero. Unlike nominal segmentation, however, targets are continuous-valued 'flux'[2] fields rather than categorical masks, with spatial coherence essential to preserve source shape;

[*] nimaseda@uw.edu
[2] A proxy of the apparent brightness of the object.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

in the original formulation, pixel values encode the exact flux of an ideally subtracted supernova, making this a spatiotemporal regression problem.

Unfortunately, building pixel–level ground truth in real sky images is far harder than in ordinary vision tasks. Truth catalogs still miss a considerable fraction of true events Mahabal et al. [2019], Sedaghat [2023] and every omission is doubly harmful: it deprives the model of clean supervision and, more importantly, may hide a genuine, perhaps unique, astrophysical event. We treat these omissions as structured label noise—systematic false negatives—and ask whether the network itself can help us recover them. While a model overfits to a sample it leaves behind a characteristic temporal fingerprint. By monitoring this "evolution cube" we show that the presence or absence of a real object can be inferred with high confidence, allowing us to retroactively fix the labels and rescue the missed science.

Learning with noisy labels (LNL) has been widely studied Wei et al. [2021], Song et al. [2022]. Most works either harden the loss against noise or identify and discard suspicious samples Frénay and Verleysen [2014], Algan and Ulusoy [2019]. Robust algorithms rely on regularization, loss correction, or robust optimisation Patrini et al. [2016], Zhang and Sabuncu [2018], whereas noise–detection approaches locate erroneous instances for subsequent cleansing Brodley and Friedl [1999]. In astronomy, dropping dubious examples is not an option: each datum may be a previously unseen phenomenon, hence our focus on precise correction rather than rejection.

Existing LNL studies concentrate on categorical labels; far fewer tackle dense, image–level regression. Exceptions explore linear regression under self–distillation Das and Sanghavi [2023] or tabular data via gradient boosting Ponti et al. [2022], but none address high-resolution astronomical imagery.

Several methods exploit training dynamics—e.g. early stopping Li et al. [2019], Liu et al. [2020], curriculum learning with dual networks Jiang et al. [2017], Han et al. [2018], or uncertainty curves Köhler et al. [2019]—yet they typically assume from-scratch training and stop at detection. We remain network–state agnostic, operate on an already competent model, and, crucially, output corrected, not merely flagged, labels. The procedure can therefore be looped: cleaned labels improve the detector, which in turn uncovers fainter events, converging towards full recovery.

Dataset Cartography Swayamdipta et al. [2020] comes closest conceptually, mapping confidence and variability of NLP classifiers to spot ambiguous samples. We adapt the spirit of that idea to pixel-level regression, start from a pre-trained astrophysical model, and emphasize the scientific imperative of preserving every single transient.

2 Problem formulation

Assume our dataset consists of two parts: a small 'gold subset' with clean labels, $\mathcal{G}$, and a larger main subset, $\mathcal{D}$, with possibly noised labels. For problem formulation, we proceed with $\mathcal{D}$ alone and come back to $\mathcal{G}$ for elaboration of the method in the next section.

Let $\mathcal{D} = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^N$ denote the dataset, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th input feature vector and $\tilde{y}_i \in \mathcal{Y}$ is the corresponding noisy label. In this paper, the label space $\mathcal{Y}$ is kept as flexible as possible. $\tilde{y}_i$ is a sample from the noisy labels, which may not reflect the true underlying labels y_i^*.

Label noise is often represented by $P(\tilde{y}_i \mid y_i^*)$: the probability of observing $\tilde{y}_i$ given the true label y_i^*. A common model is the symmetric noise model where $P(\tilde{y}_i = y_i^* \mid y_i^*) = 1 - \eta$ and $P(\tilde{y}_i \neq y_i^* \mid y_i^*) = \frac{\eta}{C-1}$ for all $\tilde{y}_i \neq y_i^*$, with $\eta \in [0, 1)$ representing the noise level. To keep the formulation as general as possible, we follow the same "instance-independent" formulation of label noise without any further assumptions, even though in the experiments we showcase the applicability of our method on datasets with more specific types of label noise too.

Let $f(\mathbf{x}; \theta)$ be the deep neural network model parameterized by θ, which maps an input $\mathbf{x}$ to an output $\hat{y} = f(\mathbf{x}; \theta)$. The loss function, which can be applied to the entire training dataset or individual mini-batches, is defined as follows:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(\mathbf{x}_i; \theta), \tilde{y}_i),$$

where $\ell(\hat{y}, \tilde{y})$ denotes a chosen loss function that measures the discrepancy between the predicted label $\hat{y}$ and the noisy label $\tilde{y}$, and N represents the total count of samples in the dataset or mini-batch.

3 Selfish Evolution: the method

Step 1: Initial training The model $f(\mathbf{x};\theta)$ is trained on the main, noisy dataset $\mathcal{D}$ for an arbitrary number of epochs:

$$\theta = \arg\min_{\theta} \mathcal{L}(\theta; x, \tilde{y}).$$

Let us refer to the intermittent state of the network after the interruption as $\dot{\theta}$

Step 2: Overfitting and evolution Consider a single-sample training dataset $\delta_i = \{(x_i, \tilde{y_i})\}$. If the goal of the model, parameterized by θ, is to minimize the loss function $L(\theta)$, overfitting occurs when $L(\theta)$ is minimized to a very small value specifically for dataset δ_i, potentially at the expense of the model's performance on other data samples. Mathematically, this can be expressed as:

$$\hat{\theta} = \arg\min_{\theta} L(\theta; x, y)$$

We let the model continue training on individual samples to overfit, capturing the evolution dynamics:

$$\theta_i^t = \text{update}(\theta_i^{t-1}, \mathbf{x}_i, \tilde{y}_i), \quad t = 1, \ldots, T,$$

where T is the number of overfitting steps and θ_i^t the model parameters at step t for sample i. We can now form the spatiotemporal "evolution cubes" which capture the changes in the model parameters and outputs during the overfitting process:

$$\mathcal{E}_i = \{\theta_i^t\}_{t=1}^T.$$

As we will show in the experiments, one can use a more generalized version of this step, where the overfitting target is not just a single sample, but a whole mini-batch, or a combination of them.

Step 3: Training of the Evolution-to-Label model We train a secondary network $g(\mathcal{E};\phi)$ parameterized by ϕ on these evolution cubes to detect and correct corrupted labels:

$$\phi^* = \arg\min_{\phi} \frac{1}{N} \sum_{i=1}^{N} \ell(g(\mathcal{E}_i;\phi), y_i^*),$$

where y_i^* are the true labels (or high-confidence corrected labels).

Closed Loop Correction One can use the trained secondary network to correct labels and iterate the process in a closed-loop fashion, aiming for a mostly clean dataset (k is the iteration index):

$$\tilde{y}_i^{(k+1)} = g(\mathcal{E}_i; \phi^{(k)}),$$

Table 1: Solver parameters used for the two parts of the evolution.

Hyperparameter	**Support**	**Selfish**
Learning Rate (α)	1e-4	1e-4
Weight Decay	0.1	0
β_1	0.99	0.9
β_2	0.999	0.999

4 Experiments

Data We use the LSST DESC DC2 simulated survey, a realistic multi-phenomena sky dataset spanning hundreds of deg^2 Abolfathi et al. [2021a,b]. Our cutouts comprise 3712 images of size 256×256 centered on 373 unique supernovae, intentionally small to emulate label-scarce astrophysical settings. We split by object to avoid leakage, yielding 3205 *train* images and 507 *gold* samples. We also construct label-noised variants at 20%, 50%, and 100% noise.

Initial training We adopt the original non-probabilistic encoder–decoder of Sedaghat and Mahabal [2018], trained on the training subset with ADAM Kingma and Ba [2014] at an initial learning rate of $1\text{e}{-}4$.

Evolution We induce a controlled "race" in the dynamics via mixed overfitting: first to a clean mini-batch (support), then to a single target (selfish). For each sample, we:

- initialize the main model from fixed pre-trained weights;

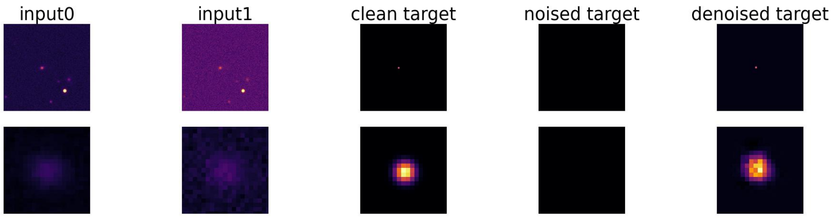

Figure 2: Denoising on an exemplar input pair. Top: full crop. Bottom: zoom on the target. “noised target” is the blank label used to train the primary network; “denoised target” is our output with the recovered ground truth.

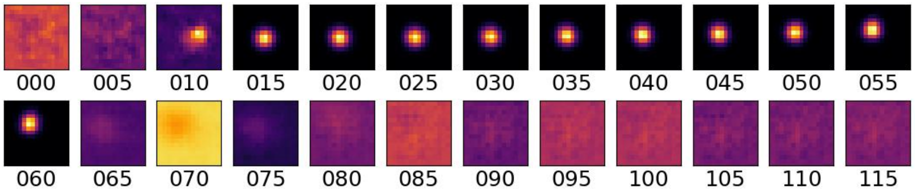

Figure 3: Down-sampled evolution cube. In the first half (top row) the network overfits a support batch; in the second, it overfits the single noisy target. The “race” between these two regimes encodes the clean label, later decoded by the E2L model.

- choose a Selfish Sample and a random clean Support Batch;
- run a short overfit on the Support Batch, appending the Selfish Sample’s per-epoch outputs to an evolution cube;
- switch to overfitting the Selfish Sample, continuing to append per-epoch outputs;
- reinitialize to the same pre-trained weights and repeat for the next sample.

Each epoch is a single iteration (one mini-batch); support-phase inference adds one forward pass on the Selfish Sample. We use ADAM with phase-specific settings in table 1 to gently bias gradients during support tuning and to allow stronger fitting in the selfish phase Mohammadi et al. [2020], Keskar and Socher [2017]. We set the number of evolution epochs $N_e = 60$.

Denoising Denoising is end-to-end: a secondary network maps evolution cubes directly to clean labels (E2L). Inputs share the main model’s tensor shape (deeper along time), and outputs are identical in type; we therefore reuse the same architecture, noting that sequence models are a drop-in alternative. To mitigate overfitting on the limited gold cubes, E2L is a thinner TransiNet (half channels) with flips (50% per axis) and shifts (uniform 0–20 px per axis). E2L trains on gold cubes from both clean and 100%-noised versions to diversify dynamics.

Table 2: Noise correction quantitative results — supernova detection

Variant	Init. Clean (%)	Cosine Sim. (%)	Hard Sim. (%)	Discovered
Baseline (full)	20.0	68.1	73.0	–
Selfish Evolution (full)	20.0	75.6	82.7	–
Baseline (500)	50.0	8.9	0.0	0
Selfish Evolution (500)	50.0	13.4	8.4	10
Baseline (full)	50.0	9.1	4.7	7
Selfish Evolution (full)	50.0	31.8	50.1	817

Results We infer corrected labels by passing evolution cubes through E2L and report: (i) cosine similarity (soft), (ii) thresholded similarity (hard), and (iii) discovery rate/count (objects recovered above threshold). Because the objective is *label correction*, evaluation is performed on the training set rather than a clean validation split. Table 2 summarizes representative runs across noise levels and data regimes: Baseline (primary network on noisy labels, no correction), “500” (only the first 500 training samples), and "full" (all training samples). In total, 817 previously missed supernovae are recovered. Figures 2 and 3 show an exemplar evolution cube and its corrected label.

5 Discussion and future work

We introduced the novel idea of detecting and correcting noisy labels in astrophysical applications based on overfitting dynamics. The proposed method helped us recover (discover) more than fifty percent of the missed supernovae in an exemplar dataset, which is beyond significant in the field of astronomy. Although we were focused on the specific task, we showcased the efficiency of the method on a rather typical classification task on MNIST, CIFAR Krizhevsky and Hinton [2009] datasets – results removed from the main manuscript due to space limitations.

Acknowledgments

AM is supported in part by NASA Grant 80NSSC24M0020.

References

Bela Abolfathi, David Alonso, Robert Armstrong, Éric Aubourg, Humna Awan, Yadu N Babuji, Franz Erik Bauer, Rachel Bean, George Beckett, Rahul Biswas, et al. The lsst desc dc2 simulated sky survey. *The Astrophysical Journal Supplement Series*, 253(1):31, 2021a.

Bela Abolfathi, Robert Armstrong, Humna Awan, Yadu N Babuji, Franz Erik Bauer, George Beckett, Rahul Biswas, Joanne R Bogart, Dominique Boutigny, Kyle Chard, et al. Desc dc2 data release note. *arXiv preprint arXiv:2101.04855*, 2021b.

Gorkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowl. Based Syst.*, 215:106771, 2019.

Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. *ArXiv*, abs/2301.13304, 2023.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25:845–869, 2014.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Wai-Hung Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Neural Information Processing Systems*, 2018.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2017.

Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jan M Köhler, Maximilian Autenrieth, and William H Beluch. Uncertainty based detection and relabeling of noisy image labels. In *CVPR workshops*, pages 33–37, 2019.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *ArXiv*, abs/1903.11680, 2019.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Ashish Mahabal, Umaa Rebbapragada, Richard Walters, Frank J Masci, Nadejda Blagorodnova, Jan van Roestel, Quan-Zhi Ye, Rahul Biswas, Kevin Burdge, Chan-Kao Chang, et al. Machine learning for the zwicky transient facility. *Publications of the Astronomical Society of the Pacific*, 131(997): 038002, 2019.

M Mohammadi, A Mohammadpour, and H Ogata. Towards theoretically understanding why sgd. In *Proceedings of the Conference on Neural Information Processing Systems*, 2020.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2233–2241, 2016.

Moacir Antonelli Ponti, Lucas de Angelis Oliveira, Juan Mart'in Rom'an, and Luis Argerich. Improving data quality with training dynamics of gradient boosting decision trees. *ArXiv*, abs/2210.11327, 2022.

Nima Sedaghat. Deep learning approach to real-bogus classification for lsst alert production. DMTN 216, LSST Data Management, Jul 2023. URL `https://dmtn-216.lsst.io/`. Version 2023-07-10.

Nima Sedaghat and Ashish Mahabal. Effective image differencing with convolutional neural networks for real-time transient hunting. *Monthly Notices of the Royal Astronomical Society*, 476(4):5365–5376, 2018.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 2022.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Conference on Empirical Methods in Natural Language Processing*, 2020.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *ArXiv*, abs/2110.12088, 2021.

Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pages 8778–8788, 2018.