Performance Comparison of Unweighted, V1-Weighted, and V2-Weighted Losses for MLP Prediction in Conway's Game of Life

Zhang Yiran

1 Weighted Loss Comparison

1.1 Overview

MLP classifiers trained with standard binary cross-entropy achieve high accuracy on Conway's Game of Life but severely under-detect the minority *alive* class. We compare three loss functions designed to address this imbalance:

- **Unweighted BCE**: $w_{\text{pos}} = 1$
- **V1 Weighted BCE**: $w_{\text{pos}} = \min(5, \sqrt{r})$
- **V2 Weighted BCE**: $w_{\text{pos}} = \min(10, r)$

where $r = N_{\text{dead}}/N_{\text{alive}}$. All models were trained across $p \in \{0.2, 0.4, 0.6\}$, regimes (early/mid/late), and patch sizes (3×3, 5×5), with 3 random seeds averaged.

1.2 Global Results

Table 1: Global averages across all densities, regimes, and patch sizes.

Model	Accuracy	Precision	Recall	F1
Unweighted	0.9167	0.6368	0.4743	0.5409
V1 Weighted	0.8950	0.4971	0.7887	**0.6092**
V2 Weighted	0.8549	0.4115	0.9060	0.5653

Summary:

- Unweighted: highest accuracy/precision, lowest recall/F1.
- V2: maximizes recall (~ 0.91) but suffers in precision.
- **V1: best F1**, achieving the most balanced trade-off.

1.3 Patch-Level Summary

Table 2: Averaged over densities and regimes.

Patch	Model	Accuracy	Precision	Recall	F1
3×3	Unweighted	**0.912**	**0.607**	0.441	0.507
	V1 Weighted	0.891	0.486	0.771	**0.595**
	V2 Weighted	0.848	0.398	**0.894**	0.550
5×5	Unweighted	**0.922**	**0.666**	0.508	0.574
	V1 Weighted	0.898	0.509	0.806	**0.623**
	V2 Weighted	0.862	0.425	**0.918**	0.581

Observations:

- 5×5 patches consistently outperform 3×3.
- V1 has the best F1 for both patch sizes.
- Best single configuration overall: **V1-weighted, 5×5 patch**.

1.4 Conclusion

Weighted losses substantially improve detection of alive cells by boosting recall. **V2** yields the highest recall but at large precision cost. **V1** delivers the strongest *balanced* performance and the highest F1, making it the preferred loss formulation for this task.

Future models (e.g. GNNs or autoregressive rollouts) should adopt V1 weighting as the default baseline.

2 Autoregressive Rollout

Setup 128×128 toroidal board; regimes/densities/burn-in as before (early: 10/40, mid: 60/40, late: 160/40); V1-weighted MLPs (3×3, 5×5), inputs are center-removed (8/24 dims), threshold 0.5.

Alignment caveat Curves below align the model's first prediction from t_0 to the true $t{+}1$ (to show drift vs. real GoL), while the training target was the current frame. A self-consistency metric (t_0 vs. first prediction) is included in the table. A full rerun with t_0-aligned accuracy is pending.

Quantitative summary (from `rollout_batch_metrics.csv`)

Patch	Init dens	Final pred dens	Final true dens	$\mathrm{Acc}_{\mathrm{self}}$	$\mathrm{Acc}_{\mathrm{first}}$	$\mathrm{Acc}_{\mathrm{last}}$
3×3	0.171	0.933	0.100	0.801	0.801	0.154
5×5	0.171	0.517	0.100	0.822	0.796	0.487

Notes: $\mathrm{Acc}_{\mathrm{self}}$ compares the first prediction to t_0 (reconstruction). $\mathrm{Acc}_{\mathrm{first}}$/$\mathrm{Acc}_{\mathrm{last}}$ compare to true $t{+}1$/ $t{+}50$. No all-zero/one collapse observed within 50 steps.

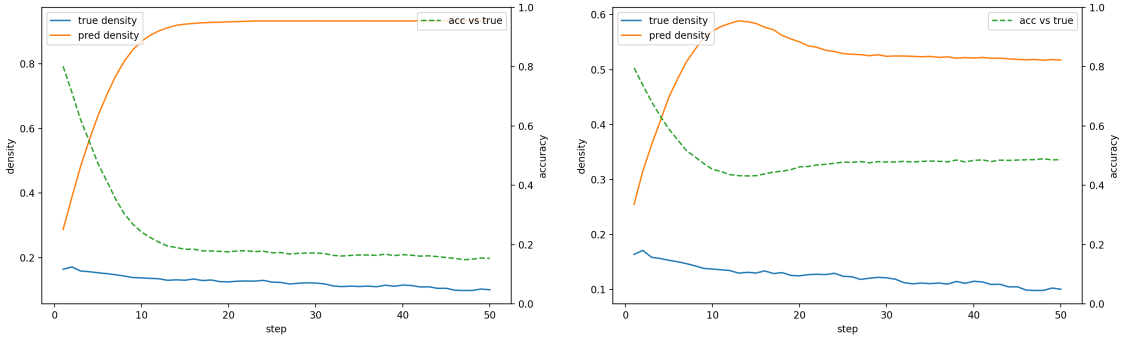

Figure 1: Density (true vs. predicted) and accuracy vs. true GoL (early, p=0.2, seed 0). Left: 3×3 quickly blows up density; right: 5×5 drifts slower.

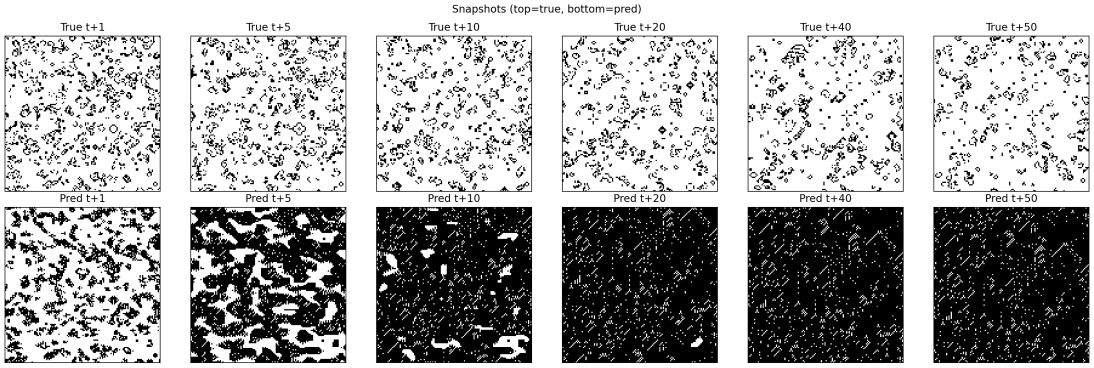

Figure 2: Snapshots (top=true, bottom=pred) at steps 1/5/10/20/40/50 for 3×3, early, p=0.2, seed 0; shows rapid over-activation.

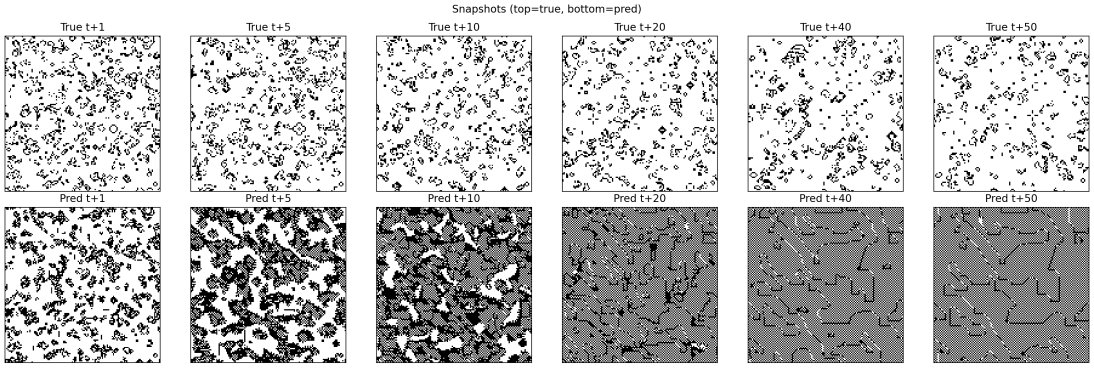

Figure 3: Snapshots (top=true, bottom=pred) at steps 1/5/10/20/40/50 for 5×5, early, p=0.2, seed 0; shows rapid over-activation.

Takeaways 5×5 shows better self-consistency and slower drift; 3×3 rapidly overpredicts alive cells and density blows up. These results illustrate divergence from true GoL under the current t+1 alignment; t_0-aligned reruns are pending.