The Reversal Ratio Is an Optimisation Artefact

Matched-Step Evaluation of Synthetic Data Augmentation

Vladimir Nelyub · Vadim Tynchenko · Aleksei Borodulin · Ahmad Hammoud · Sergei Kurashkin

Bauman Moscow State Technical University, Moscow, Russia · kurashkin@bmstu.ru

CDEL: 2nd Workshop on Curated Data for Efficient Learning · ECCV 2026, Malmö

ECCV 2026 | MALMÖ | SEPT 8–12

The question

Generative augmentation is measured almost exclusively where real data is plentiful, and there added data tends to help monotonically. We measure the reversal ratio r*, the synthetic-to-real ratio at which added synthetic data stops helping, in two narrow domains and a broad natural-image control.

At a fixed epoch count a larger training set also buys proportionally more gradient steps.

The design

3 × 2
domains × real-data budgets

2
generator provenance legs

0.25–8
synthetic-to-real ratios

5
classifier seeds per cell

1,209
gated units (609 + 600)

23.19
GPU-hours, main grid

One fixed classifier recipe and one fixed generator configuration per leg apply to every condition, so the fairness argument is an identical budget rather than tuned optima and no optimism gap arises.

Breadth, not quantity

Domain	Classes	Scarce	Moderate
EuroSAT narrow · satellite	10	500 (50)	2000 (200)
TrashCan 1.0 narrow · underwater	9	450 (50)	1800 (200)
CIFAR-100 broad control	100	2500 (25)	10000 (100)

Per class the control is the poorer setting, so the domain effect cannot be read as the control simply holding more data per category. The domains also differ in class count, native resolution, and similarity to the external generator's source corpus; these vary together and we do not separate them.

What the gate checks

Every number comes from one gated pass with resume disabled. The gate checks the clean run, external validity across the three public datasets, the presence of the statistics files, and both frozen generator legs in every cell.

Protocol

Real set R
class-balanced, two budgets per domain

scarce-adapted G
StyleGAN2-ADA trained on the target's own R only

external-pretrained G
an FFHQ source network fine-tuned to the target

one generator frozen per (domain, budget, leg), sha256-hashed, with an external-data flag

Synthetic set S, |S| = r |R|
r = 0.25, 0.5, 1, 2, 4, 8

ResNet-18 from scratch on real + synthetic
40 epochs, SGD, cosine schedule, batch 128

Clean test and five corruptions at severity 3
acc_shift is the primary metric

Anchors: real-only, count-matched classical augmentation, and a real-data oracle.

The reversal ratio r*

r* is the grid argmax of the mean acc_shift curve over r = 0, 0.25, 0.5, 1, 2, 4, 8, with a bootstrap 95% interval over five seeds.

r* = 0: no positive ratio improves on real-only.

r* ≥ 8: right-censored. No reversal occurs inside the grid, so the cell bounds the value from below and is never read as a reversal at 8.

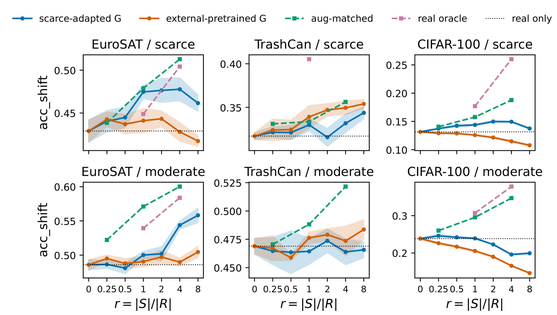
Two provenance legs

One architecture family, two histories: the scarce-adapted leg sees only the target's own real data, while the external-pretrained leg transfers a synthesis trunk that has seen an external corpus.

External pretraining helps on TrashCan, is at parity on EuroSAT, and costs 0.092 accuracy at r = 8 on CIFAR-100/moderate.

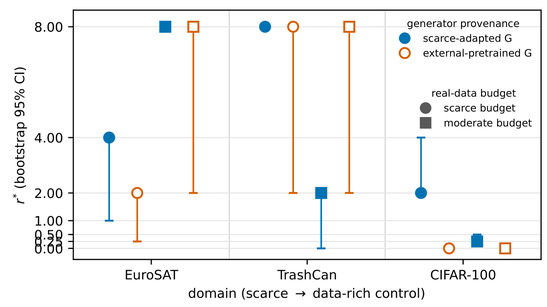
Disclosed confound: the legs train for 400 and 150 kimg, so the external leg's deficit could reflect less adaptation rather than the wrong prior. We report provenance as a confound rather than as an effect size.

Scaling curves, fixed epochs



acc_shift against r, per domain (columns) and real-data budget (rows). Synthetic data keeps helping to large r in the narrow domains and reverses almost immediately in the broad control, where the external-pretrained leg falls fastest.

Reversal ratio, fixed epochs



Grid argmax with bootstrap 95% intervals; shape encodes the budget and fill encodes provenance. Points at the top of the grid are right-censored: five of the eight narrow-domain cells. In the control r* falls to 0.25, and to 0 on the external leg at both budgets.

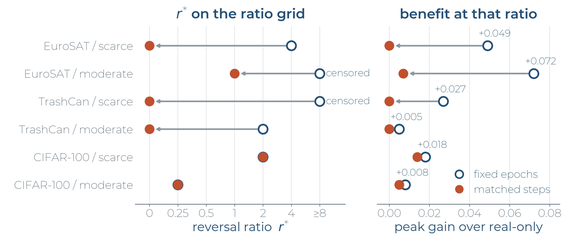
Robustness, not clean accuracy

On clean accuracy r* collapses to 0, or at most 2, almost everywhere, while on acc_shift the narrow domains retain r* up to 8. Synthetic data extends the useful ratio under corruption shift well past the point where it stops helping on clean images.

Synthesis stays below the real-data oracle in eight of nine comparable cells, and the oracle gap widens with r and with domain richness, reaching −0.212 at CIFAR-100/moderate and r = 4.

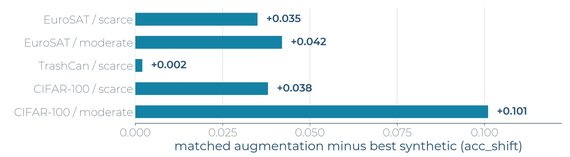
The Friedman omnibus over the eighteen conditions is significant (Iman–Davenport $F(17,51) = 7.92$, $p = 3.9e{-9}$), but with four complete blocks the Nemenyi critical difference is wide, so only the extremes of the ranking separate.

Matched steps remove it



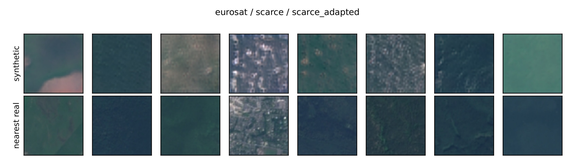
Scarce-adapted leg, step budget pinned to real-only. No cell stays censored.

Classical augmentation wins



Under matched steps augmentation retains up to +0.077 against at most +0.015 for synthesis, and the two do not combine.

They do not memorise



Synthetic samples above their nearest real training image. The weakness is elsewhere: recall never exceeds 0.14, and label fidelity on the hundred-class control is 0.01, at chance.

What we recommend

- Match the optimisation budget as well as the sample count before attributing a gain to added data.
- Report r* against an augmentation-matched baseline and a real-data oracle.
- Freeze the generator, hash it, and record whether external data was seen.
- Audit coverage and class conditioning alongside sample fidelity.

Limitations: one generator family, a corruption proxy for natural shift, a classifier trained from scratch, and one generator seed per cell, so the intervals carry classifier variance only.

Code, per-unit results, manifests and the twelve generator hashes: doi.org/10.5281/zenodo.21928051

code and data