

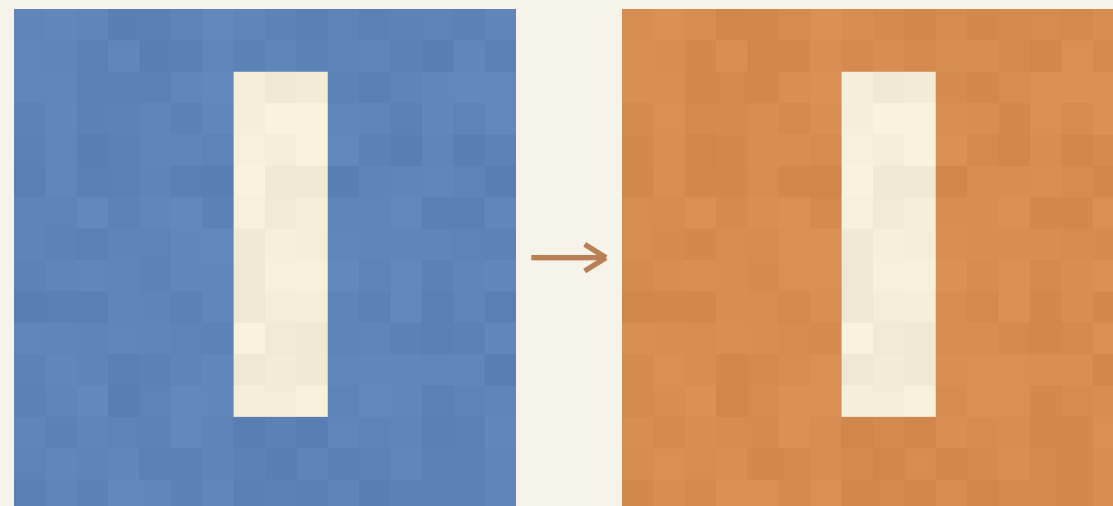
See the shortcut.

Did the model learn the shape, or the backdrop?

A working visual ML experiment. Train a real classifier, change one irrelevant factor, then repair the training lesson.

One shape. Two answers.

Controlled synthetic data. Real optimization.



UPRIGHT

Model: upright

72.4% model score

UPRIGHT

Model: crossbar

69.3% model score

The same held-out specimen; only its backdrop changes.

A perfect score can hide a bad rule.

ORIGINAL TEST WORLD

100%

192 / 192 held-out tiles

Every paired prediction flips. The original score alone concealed the model's reliance on an irrelevant visual cue.

SWAP ONLY THE BACKDROP

0%

0 / 192 paired counterfactuals

Seed 42. Geometry, foreground, texture and true label stay fixed. These are toy-task results, not a field benchmark.

Change the lesson, not the answer.

01

GENERATE

128 original training specimens. Two shapes. Irrelevant color correlated with label.

02

FIT

768 pixel features. Regularized logistic regression. 120 actual gradient updates.

03

INTERVENE

192 separate held-out specimens. Swap only the backdrop; compare each pair.

04

RETRAIN

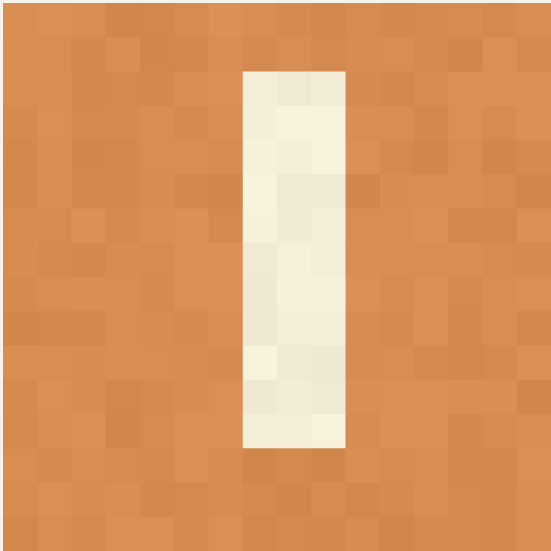
Pair both backdrops for each training geometry. Start again from zero.

Held-out examples never enter training. Inference can read pixels, never labels or backdrop metadata.

The answer survives the intervention.

PAIRED BACKGROUND TRAINING	BEFORE	AFTER
Original-world accuracy	100%	100%
Swapped-backdrop accuracy	0%	100%
Worst of four groups	0%	100%
Prediction flips	100%	0%

Small prediction margins remain. Perfect observed accuracy is not certainty or general robustness.



Model: upright

53.3% uncalibrated score

The experiment is open all the way down.

Original data

Procedural 16 x 16 RGB tiles. Disjoint geometry-seed domains for training and held-out tests.

Inspectable learning

Explicit loss, gradient and optimizer. Numerical gradient checks validate the implementation.

Meaningful checks

15 passing local tests: pixel invariance, leakage boundaries, seed variation, malformed imports and reproducible weights.

Portable evidence

Export a small recipe. Replay regenerates the data and retrains both models; saved summary values are never trusted.

A lesson people can test.

Next: test whether learners can spot shortcuts in a different task after using this workbench.

No learner study or real-world model audit is claimed. Extend to other controlled factors only while preserving explicit data separation and uncertainty.

toukoursin.github.io/shiftglass

github.com/ToukoUrsin/shiftglass



MIT source and original examples. No account, external inference, credentials or private data required.

Substantial AI assistance in coding, design, tests and writing is disclosed. Runtime: the implemented local classifier.