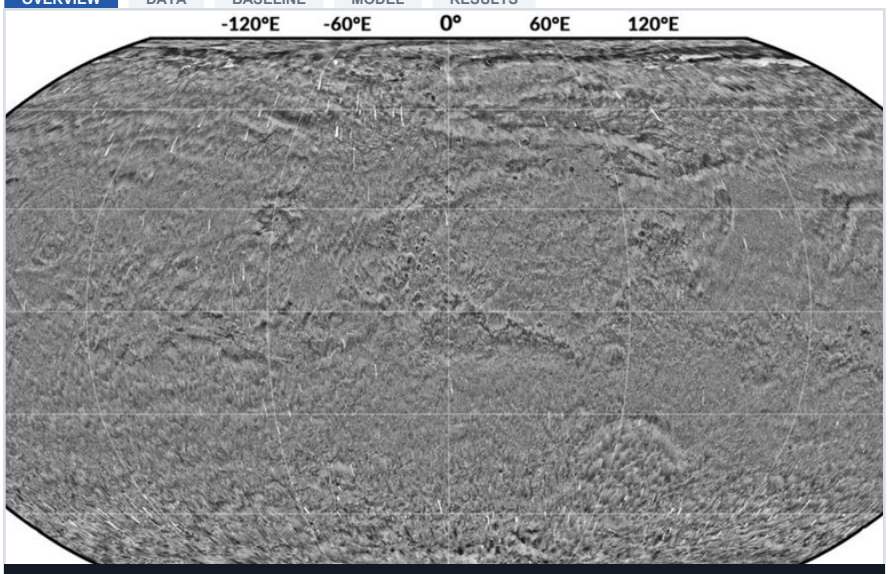
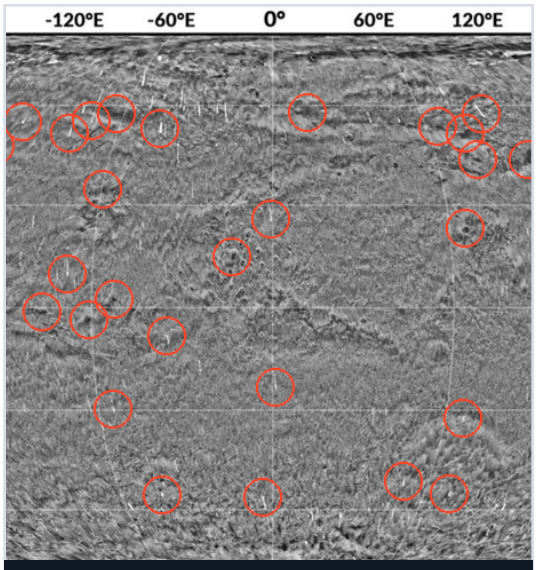
MARS CTX CRATER DETECTION

FROM NOTEBOOK RECONSTRUCTION TO A LOCKED, CONTROLLED FASTER R-CNN TEST

OVERVIEW DATA BASELINE MODEL RESULTS



OFFICIAL CTX MOSAIC BROWSE IMAGE



CANDIDATE-GENERATOR SMOKE TEST

WHAT THE PROJECT ACTUALLY BECAME

The work began as a reconstruction of a shared Mars crater notebook. It matured into an evidence-first benchmark: provenance-controlled CTX imagery, catalog-to-tile labels, geographic leakage guards, one-to-one evaluation, an intentionally weak YOLO baseline, and a later isolated Faster R-CNN experiment with a locked validation decision and one final test evaluation.

PIPELINE

CTX GeoTIFF > 1024 px tiles > Catalog boxes > Faster R-CNN > 1:1 IoU match > GIS output

MY TECHNICAL CONTRIBUTION

- Built the evaluation and evidence layer around the preserved notebooks instead of overwriting the original research files.
- Implemented catalog audits, affine label generation, geographic splits, map-level deduplication, threshold sweeps, and traceable run manifests.
- Designed and executed the isolated S2 Experiment 3 comparison, preserving selection/test separation and publishing failure cases alongside headline metrics.

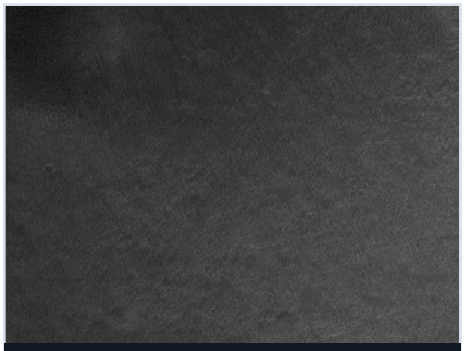
CLAIM BOUNDARY

The final result is a measured small-data detector baseline, not a production crater catalog. Weak baselines and missing detections are retained because they explain what the next experiment must fix.

2,809	1,683	1
TOTAL GENERATED TILES	CATALOG CRATERS IN BENCHMARK	LOCKED TEST EVALUATION

DATA, LABELS, AND LEAKAGE CONTROL

THE BENCHMARK WAS ENGINEERED BEFORE THE MODEL WAS TRUSTED

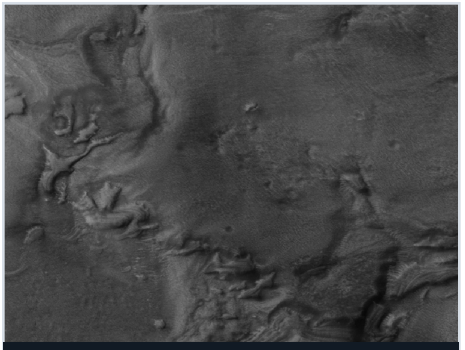

GROUND TRUTH TILE 1


GROUND TRUTH TILE 2


GROUND TRUTH TILE 3

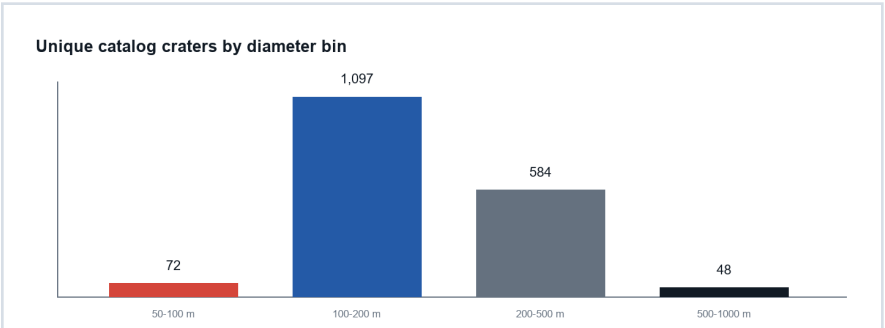

GROUND TRUTH TILE 4


GROUND TRUTH TILE 5


GROUND TRUTH TILE 6

SOURCED DATASET CONSTRUCTION

The E028_N40 CTX region was cut into overlapping 1024 x 1024 tiles with a 128 px overlap and 896 px step. Catalog crater centers and DIAM_M values were projected through raster metadata to form tile-local boxes. Diameter in pixels was derived from trusted meters-per-pixel metadata, never from screenshot scale.



CATALOG DIAMETER DISTRIBUTION

LEAKAGE GUARD

A deterministic geographic-block split kept nearby overlapping tiles and repeated crater IDs out of different splits. The audit found zero cross-split overlapping tile pairs and zero crater IDs present in multiple splits.

DATASET AUDIT

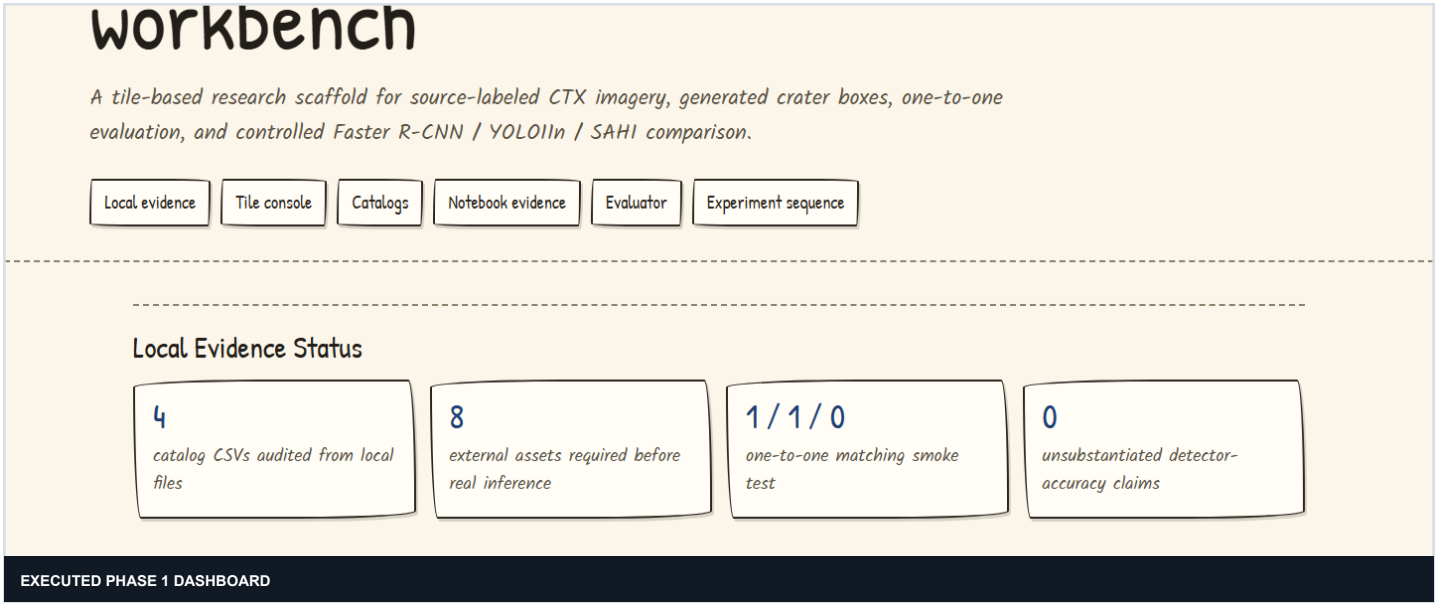
1,908 train tiles | 318 validation | 371 test | 212 excluded by guard. The 2,344 label instances represent 2,004 complete and 340 clipped boxes; 1,561 tiles are negatives.

0	75-930 m	1,561
CROSS-SPLIT OVERLAPS	CATALOG DIAMETER RANGE	NEGATIVE TILES

THE NECESSARY FAILED BASELINE

A ONE-EPOCH YOLO11N RUN PROVED THE PIPELINE AND FAILED THE SCIENCE TARGET

OVERVIEW | DATA | **BASELINE** | MODEL | RESULTS



EXECUTED PHASE 1 DASHBOARD

WHY RUN A DELIBERATELY SMALL BASELINE?

The first local run was designed to prove that sourcing, tiling, training, inference, evaluation, and map-level deduplication could execute end to end on real CTX data. It used one YOLO11n epoch. Its purpose was instrumentation, not accuracy.

MEASURED OUTCOME

- 371 test tiles; 15.71 s inference; 42.3 ms mean per tile.
- 111,300 raw detections became 53,878 after NMS,
- overlap merge, and regional GIS deduplication.
- At IoU 0.5: 0 TP, 53,878 FP, 143 FN; AP50 = 0.0.
- Validation mAP50 was 6.11e-7.

ENGINEERING INTERPRETATION

This was not hidden or renamed a success. The run showed that the complete data path worked, but the detector had learned almost nothing useful after one epoch. That evidence justified a smaller, controlled Faster R-CNN experiment with balanced sampling and explicit validation selection.

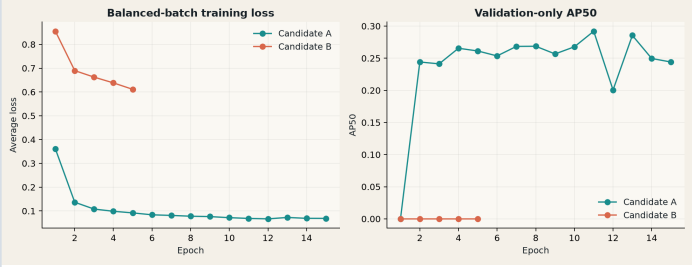
WHAT THE FAILURE RULED OUT

- Pipeline completion alone is not detector performance.
- Large candidate volume cannot substitute for one-to-one matches against catalog craters.
- Reporting speed without false-positive burden would misrepresent operational usefulness.

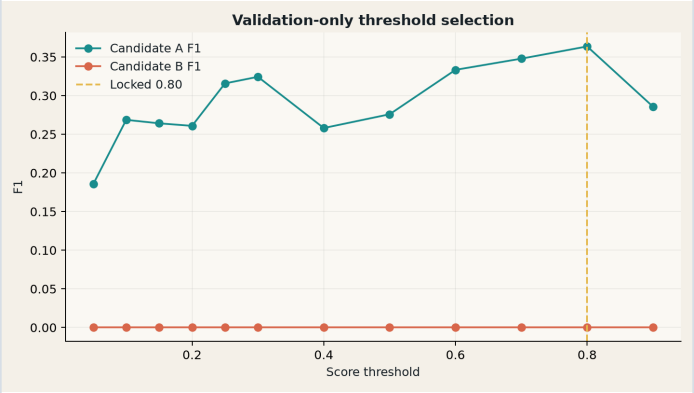
0.0	53,878	42.3 ms
TEST AP50	POST-DEDUPE FALSE POSITIVES	MEAN INFERENCE / TILE

CONTROLLED FASTER R-CNN EXPERIMENT 3

TWO CANDIDATES, VALIDATION-ONLY SELECTION, THEN ONE LOCKED TEST



CANDIDATE A AND B TRAINING HISTORIES



VALIDATION-ONLY THRESHOLD SELECTION

PROTOCOL

- Both candidates used balanced batches of four tiles: two positive and two negative, 38 steps per epoch.
- Candidate A ran 15 epochs; its best validation AP50 was 0.292 at epoch 11. Candidate B stopped after five epochs with validation AP50 = 0.
- The choice of A and score threshold 0.8 used validation data only. Frozen test inputs were prohibited from tuning and debugging.
- The locked checkpoint SHA-256 and audit bundle were recorded; 14 automated tests passed before the final evaluation.

LOCKED DECISION

Candidate A | epoch 11 | score threshold 0.8 | IoU threshold 0.5. The test set was evaluated once after the selection was frozen.

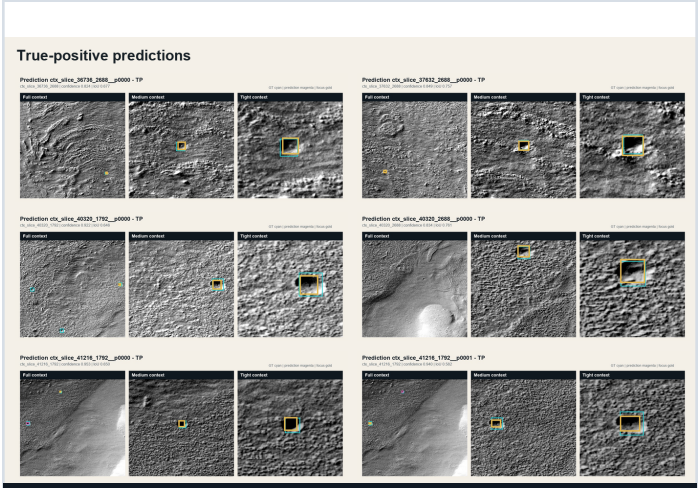
WHY CANDIDATE B MATTERED

B's zero-AP result showed that a nominal architecture variant could collapse to predicting nothing at its selected threshold. Keeping that negative result prevented a post-hoc story in which every attempted model appeared useful.

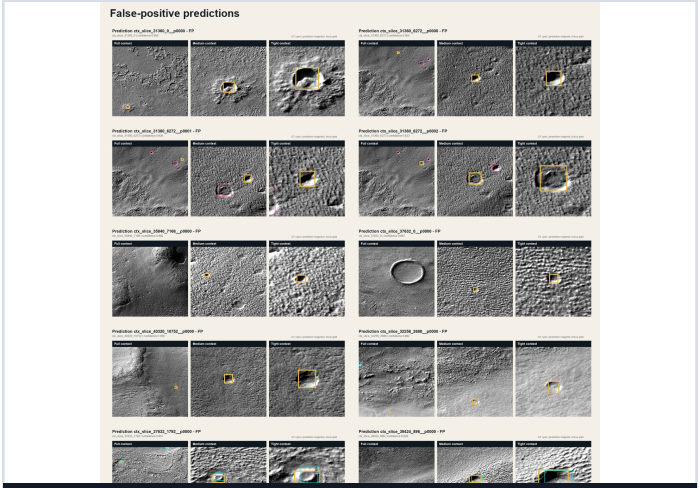
0.292	0.8	14 passed
BEST VALIDATION AP50	LOCKED SCORE THRESHOLD	PIPELINE TESTS

LOCKED TEST: RESULTS AND FAILURE MODES

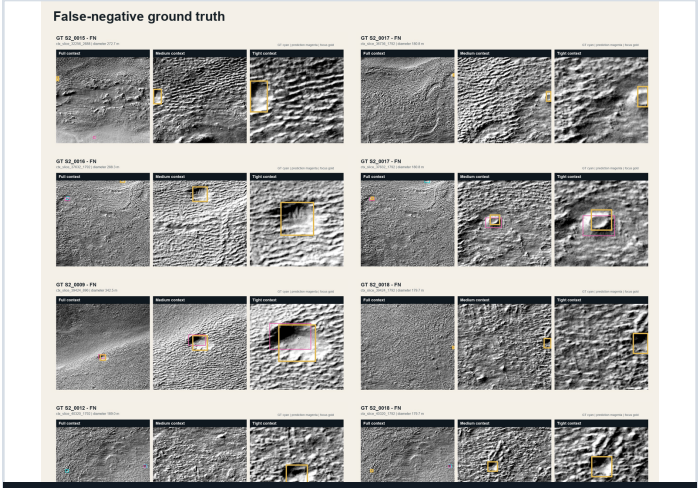
THE MODEL FOUND REAL CRATERS, BUT RECALL AND FALSE POSITIVES REMAIN LIMITING



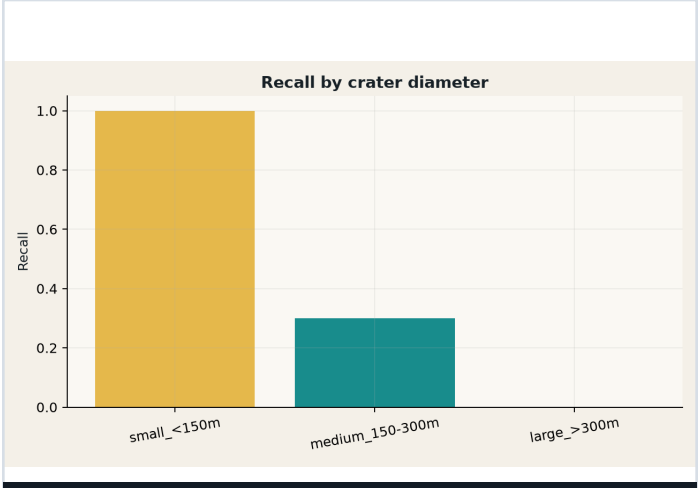
TRUE-POSITIVE PREDICTIONS



FALSE-POSITIVE PREDICTIONS



MISSED GROUND-TRUTH CRATERS



RECALL BY CRATER DIAMETER

FINAL LOCKED METRICS

At score threshold 0.8 and IoU 0.5: 6 TP, 10 FP, and 8 FN. Precision = 0.375, recall = 0.429, F1 = 0.400, AP50 = 0.263, and median matched IoU = 0.663. Seven false positives occurred on negative tiles and three on positive tiles.

WHAT THE IMAGES SAY

Small-crater recall was strongest in the sampled bins while medium and large examples remained sparse. Misses cluster in low-contrast, degraded, or terrain-confused cases; false positives often resemble circular landforms. More diverse labeled geography is a stronger next move than tuning on this test set.

NEXT EXPERIMENT

Freeze this result. Expand geographically separated training data, preserve hard negatives, compare threshold behavior on validation only, and repeat with a new untouched test set. Do not reuse this test set for selection.

0.263	0.375 / 0.429	0.663
AP50	PRECISION / RECALL	MEDIAN MATCHED IOU