From Satellite to UAV Imagery: Calibration and Collapse-Aware Triage for Trustworthy Post-Earthquake Building Damage Assessment

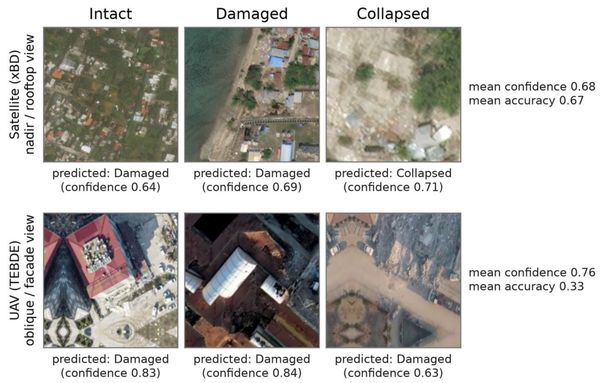
OntarioTech Engineering & Applied Science

Mustafa Eren Soyhan, Khalid Elgazzar · Department of Electrical, Computer and Software Engineering

IoT Research Lab

The problem

- After an earthquake, responders must find collapsed buildings within hours. Manual review of imagery does not scale.
- Damage classifiers are trained on satellite data such as xBD, which shows rooftops from above. Drones arrive first and see facades obliquely.
- Labels from the new event arrive late or never, so decisions ride on the model's confidence.



Same severity, two viewpoints: on UAV crops the model predicts “Damaged” with 0.76 confidence at only 0.33 accuracy.

Our approach is a trustworthiness layer

- Keep the satellite-trained classifier frozen; add two cheap components on top.
- **Target calibration:** one temperature fitted on a small UAV sample makes confidence honest.
- **Collapse-aware triage:** rank human review by predicted collapse probability, not by low confidence.

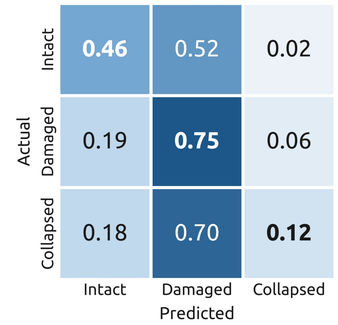
Validated on two real disaster scenarios

1,636
UAV patches, Türkiye 2023 earthquakes

10,863
UAV patches, Hurricane Michael

3 × 3
network architectures and random seeds

Zero-shot transfer fails, confidently



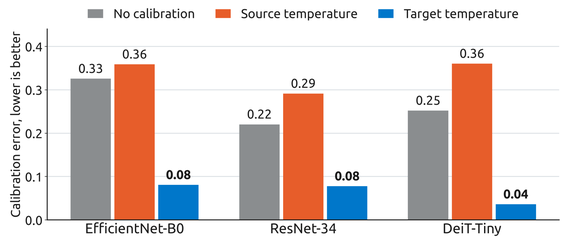
0.479
accuracy, down from 0.90 on satellite data

0.118
collapsed recall: 88% of collapses missed

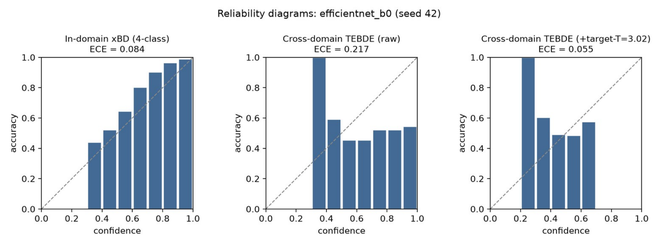
0.326
expected calibration error

EfficientNet-B0 on UAVs-TEBDE: Intact and Collapsed both flow into “Damaged”.

Calibration must come from the target



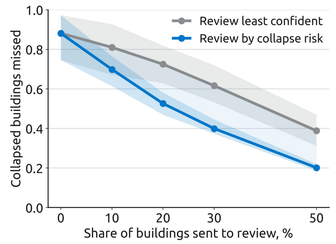
Calibration error on UAVs-TEBDE, mean over three seeds.



Reliability: in-domain, cross-domain raw, cross-domain with target temperature.

- A target temperature restores calibration error to 0.04 to 0.08 for every backbone with accuracy unchanged, $p < 0.001$. Fifty labeled patches are enough.
- Reusing the source temperature makes it worse, from 0.326 to 0.359: calibration does not transfer across the shift.

Collapse-aware triage halves the misses



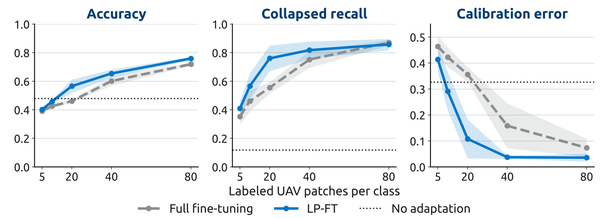
0.616 → 0.400
collapsed-miss-rate at a 30% review budget

0.389 → 0.202
collapsed-miss-rate at a 50% review budget

EfficientNet-B0 on UAVs-TEBDE; bands show the range over seeds.

- Lower at every budget, for every backbone, with no target labels, $p \le 0.0001$.
- The rule costs a division and a sort: review the buildings most likely to be collapsed, not the least confident ones.

Twenty labels go a long way



EfficientNet-B0 with a few labeled UAV patches; bands show the spread over runs.

- Label-free adaptation with BN-adapt or TENT degrades accuracy and calibration for 2.3 to 2.7× the compute.
- LP-FT with **20 labels per class**: accuracy 0.565 and collapsed recall 0.760, $p = 0.004$. Calibration error falls below 0.05 by 40 labels.

It replicates, and it is cheap

0.19 to 0.32
Hurricane Michael misses, down from 0.50 to 0.60

11 to 23 ms
inference per building patch

1 to 4 µs
added by calibration and triage

Field recipe

1. Fit one temperature on a small sample of target imagery. Never reuse the source one.
2. Skip label-free test-time adaptation under viewpoint shift.
3. Spend the first 20 to 40 labels per class on linear probing, then fine-tuning.
4. Until labels exist, rank human review by collapse probability.

Published in
IEEE UBMK 2026, IEEE Xplore
DOI 10.1109/UBMK71847.2026.11745345

Contact
mustafa.soyhan@ontariotechu.net

Scan for the paper