RescueLens AI

December 2024

Disaster Damage Assessment Methodology Recommendations

METHODOLOGY ANALYSIS: FEMA PDA Image Assessment

This document outlines recommended approaches for improving the accuracy and consistency of AI-powered disaster damage assessment that matches images to FEMA Preliminary Damage Assessment (PDA) standards.

CURRENT APPROACH ASSESSMENT

What You Are Doing:

Zero-shot LLM vision (Claude Opus) + detailed FEMA prompting + structured JSON output

Strengths:

- Flexible - handles novel damage types without retraining
- Explainable - produces detailed justifications for each classification
- No training data required to get started
- Easy to update rules (just edit the system prompt)

Weaknesses:

- Inconsistent across similar images (no memory between assessments)
- Can miss subtle visual cues (water stains, hairline cracks)
- Expensive at scale (~$0.15+ per image with Opus)
- Does not learn from corrections

RECOMMENDED METHODOLOGIES (Ranked by ROI)

1 FEW-SHOT LEARNING WITH REFERENCE IMAGES

BEST ROI

Add 2-3 labeled example images PER severity level to your prompt. This dramatically improves consistency by giving the model calibration points.

Example prompt addition:

"Here's an example of MAJOR damage: [image] - water line at outlet height"

"Here's an example of MINOR damage: [image] - water below outlets"

Why it works: LLMs are great pattern matchers but need anchors. Reference images give calibration.

Effort: Low (collect 15-20 good reference images)

Impact: High (30-50% consistency improvement)

2 HYBRID: LLM + CUSTOM CLASSIFIER

Train a lightweight CNN to detect specific features (water line height, structural breach, home type), then feed those detections TO Claude as structured input.

Example workflow:

1. CNN detects: "Water line at 24 inches"
2. Feed to Claude: "Classify this conventional home with water at 24 inches"

Why it works: CNNs excel at repetitive visual tasks. LLMs excel at reasoning. Combine both.

Effort: Medium (need ~500+ labeled images per class)

Impact: Very High

3 HUMAN-IN-THE-LOOP + PROMPT REFINEMENT

Track where the model fails and WHY. Common failure patterns include: overestimates damage from debris, misses water stains, confuses accessory structures with primary dwelling.

Then add explicit rules for each failure mode:

"CRITICAL: Tree in yard does not equal tree through roof. Verify penetration."

Why it works: Your FEMA rules are good. The issue is usually edge cases not covered.

Effort: Low (logging + prompt iteration)

Impact: Medium-High

4 FINE-TUNED VISION MODEL (Future)

If you accumulate 5,000+ labeled images, fine-tune a vision model like GPT-4V (OpenAI), Qwen-VL (open source), or Florence-2 (Microsoft).

Why it works: Domain-specific training beats general prompting for specialized tasks.

Effort: High (dataset collection, training infrastructure)

Impact: Highest (but requires scale)

RECOMMENDED ACTION PLAN

1. IMMEDIATELY:

Add 2-3 reference images per severity level to the prompt (few-shot learning)

2. SHORT-TERM:

Log assessment failures and add explicit error-correction rules to the prompt

3. MEDIUM-TERM:

Build a water-line height detector (CNN) to feed measurements to Claude

4. LONG-TERM:

Collect a labeled dataset of 5,000+ images for fine-tuning

COMPARISON SUMMARY

Method	Effort	Impact	Data Needed
Few-Shot Reference Images	Low	High	15-20 images
Hybrid LLM + CNN	Medium	Very High	500+ per class
Human-in-the-Loop	Low	Medium-High	Failure logs
Fine-Tuned Model	High	Highest	5,000+ images