🔎 Interpretation

- **Overall (all classes):**
  - Box mAP50 = **0.623**
  - Box mAP50-95 = **0.425**
  - Mask mAP50 = **0.618**
  - Mask mAP50-95 = **0.373**
- **Best Performing Class:**
  Corrosion Induced Spalling with strong precision, recall, and high mAP values (Box mAP50=0.821, Mask mAP50=0.814).
- **Weakest Class:**
  Crack detection/segmentation with low recall (0.312 Box, 0.297 Mask) and low mAP50-95 (0.233 for Box, 0.113 for Mask). This suggests cracks are harder for the model to capture.
- **Balanced Class:**
  Peeling shows very good recall (0.822) and decent precision, with high mAP50 (0.723 Box, 0.715 Mask).

Class	Images	Instances	Box(P)	Box(R)	Box(mAP50)	Box(mAP50-95)	Mask(P)	Mask(R)	Mask(mAP50)	Mask(mAP50-95)
all	400	727	0.654	0.608	0.623	0.425	0.660	0.617	0.618	0.373
Corrosion Induced Spalling	98	157	0.751	0.771	0.821	0.525	0.763	0.783	0.814	0.442
Crack	100	196	0.617	0.312	0.411	0.233	0.587	0.297	0.365	0.113
Peeling	90	163	0.671	0.822	0.723	0.592	0.671	0.822	0.715	0.570
Spalling	100	211	0.577	0.526	0.535	0.351	0.618	0.564	0.577	0.368

📑 Documentation: Understanding YOLO Segmentation Metrics

When training a YOLO (You Only Look Once) segmentation model, we evaluate its performance using several key metrics. These metrics measure how accurately the model

detects and segments defects (e.g., cracks, spalling, peeling) in images. Below is a simple explanation of each metric, tailored for engineers who may not have a background in AI.

---

1. Precision (P)

- **Definition:** Out of all the defects the model predicts, how many are actually correct.

- **Analogy (Civil Engineering):** If you mark 10 cracks on a bridge photo, and 7 are truly cracks while 3 are mistakes, your precision is 70%.
- **Formula:**

$$\text{Precision} = \frac{TP}{TP + FP}$$

Where:

- **TP (True Positives):** Correct predictions (e.g., correctly identified cracks).
- **FP (False Positives):** Wrong predictions (e.g., model marked a crack where none exists).

2. Recall (R)

- **Definition:** Out of all the real defects present, how many did the model successfully find.
- **Analogy:** If there are 10 actual cracks in a wall, and your model finds 6 of them, recall = 60%.
- **Formula:**

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where:

- **FN (False Negatives):** Missed predictions (real cracks the model failed to detect).

3. mAP50 (Mean Average Precision at IoU = 0.5)

- **Definition:** A balanced measure of accuracy that considers both precision and recall. It checks whether predicted defects overlap with actual defects by **at least 50%**.
- **Analogy:** If the model draws a box around a crack, at least half of that box must overlap with the real crack region to count as correct.

- **Formula (simplified):**

$$\text{AP} = \int_0^1 P(R)\, dR$$

(mAP = Average of AP across all classes and defects)

- **mAP50:** Uses a **50% overlap threshold**.
- **Interpretation:** Easier to achieve; more forgiving.

4. mAP50-95

- **Definition:** Same as mAP, but calculated at **multiple overlap thresholds (0.5, 0.55, …, 0.95)**.
- **Analogy:** Stricter checking — not only should the crack box overlap by 50%, but also by 75%, 90%, etc. The model must be precise at all levels.
- **Formula:**

$$\text{mAP}_{50-95} = \frac{1}{10} \sum_{t=0.5}^{0.95} \text{mAP}_t$$

Where ttt are the thresholds (0.5, 0.55, 0.6, …, 0.95).

5. Box Metrics vs Mask Metrics

- **Box Metrics:** Evaluate **bounding boxes** (rectangular areas around defects).
- **Mask Metrics:** Evaluate **segmentation masks** (pixel-accurate outlines of defects).

Example:

- **Box:** Model draws a rectangle around a crack.
- **Mask:** Model traces the actual crack shape pixel by pixel.

6. Why These Metrics Matter

- **Precision** → Reduces false alarms (e.g., wrongly marking stains as cracks).
- **Recall** → Ensures no defects are missed (important for safety).
- **mAP50** → Measures practical usefulness of detection (looser check).
- **mAP50-95** → Measures strict accuracy (closer to real inspection standards).
- **Mask metrics** → Important for quantifying **damage area** in pixels or mm², which can be linked to **structural deterioration assessment**.

Class-wise Interpretations

1. Corrosion Induced Spalling

- **Precision:** 0.751 (75.1%) → Most detections are correct.
- **Recall:** 0.771 (77.1%) → The model successfully captures the majority of spalling cases.
- **mAP:** Box = 0.821, Mask = 0.814 → High accuracy in both bounding box and mask segmentation.
- **Interpretation:** This is the **best-performing class**, with balanced precision and recall. The model is reliable for identifying corrosion-induced spalling in structures.

2. Crack

- **Precision:** 0.617 (61.7%) → Over 1/3 of crack detections are incorrect.
- **Recall:** 0.312 (31.2%) → The model misses many real cracks.
- **mAP:** Box = 0.411, Mask = 0.365 → Low detection and segmentation accuracy.
- **Interpretation:** Cracks are the **weakest class**. Low recall suggests that cracks are often too fine or irregular for the model to detect consistently. Additional data or specialized preprocessing may be needed.

3. Peeling

- **Precision:** 0.671 (67.1%)
- **Recall:** 0.822 (82.2%) → Very high recall.
- **mAP:** Box = 0.723, Mask = 0.715 → Strong overall accuracy.
- **Interpretation:** The model is very effective at **finding peeling defects**, rarely missing them. This makes it suitable for inspection tasks where peeling surfaces must not be overlooked.

4. Spalling

- **Precision:** 0.577 (57.7%) → More false detections compared to other classes.
- **Recall:** 0.526 (52.6%) → About half of spalls are detected.
- **mAP:** Box = 0.535, Mask = 0.577 → Moderate detection and segmentation quality.
- **Interpretation:** Performance is **average**, but weaker than corrosion-induced spalling or peeling. The model finds some spalls but misses others and produces false alarms.

📈 Overall Performance (All Classes)

- **Precision (Box/Mask):** ~65% → About two-thirds of detections are correct.
- **Recall (Box/Mask):** ~61% → The model detects a little over half of all real defects.
- **mAP50:** ~0.62 → Model achieves reasonable accuracy at 50% overlap threshold.
- **mAP50-95:** ~0.37–0.43 → Under stricter overlap thresholds, performance decreases, showing that predictions are sometimes imprecise in localization.

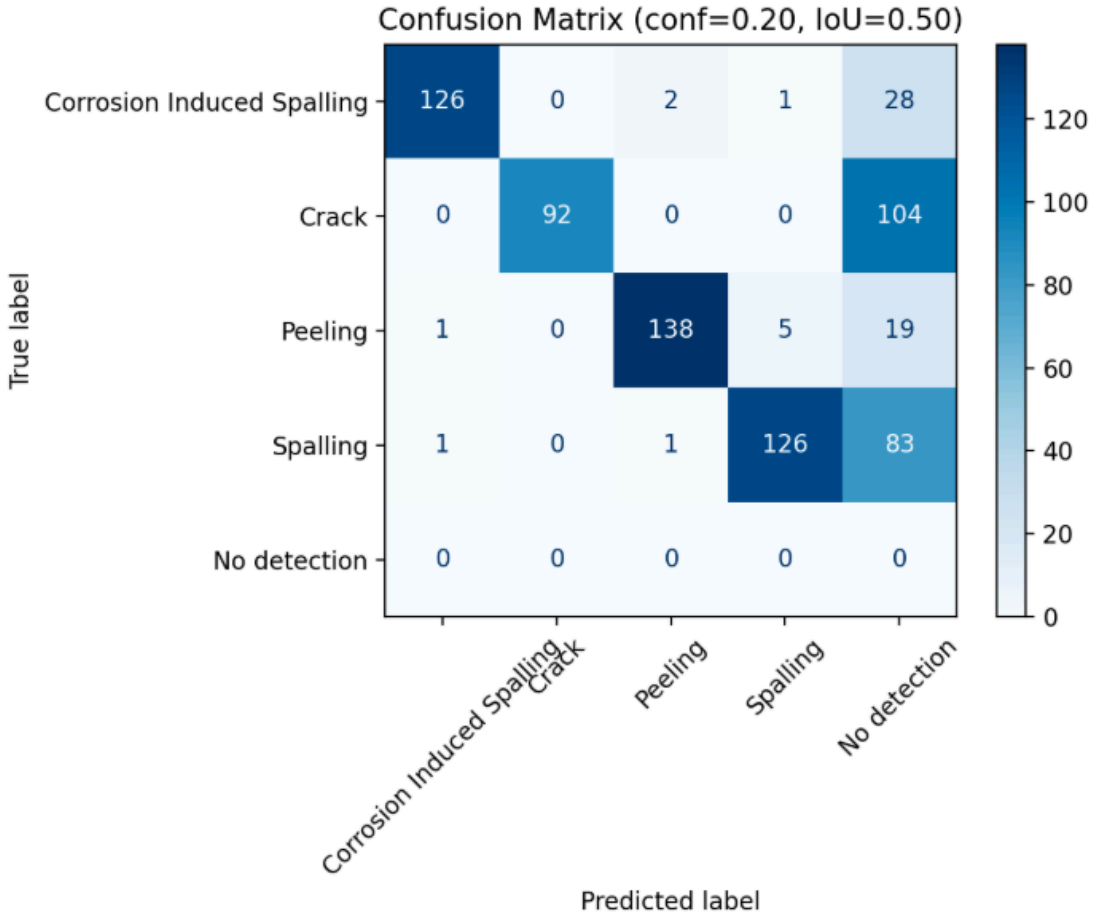
🏗️ Engineering Interpretation

- The model is **very reliable for detecting Corrosion Induced Spalling and Peeling** defects.

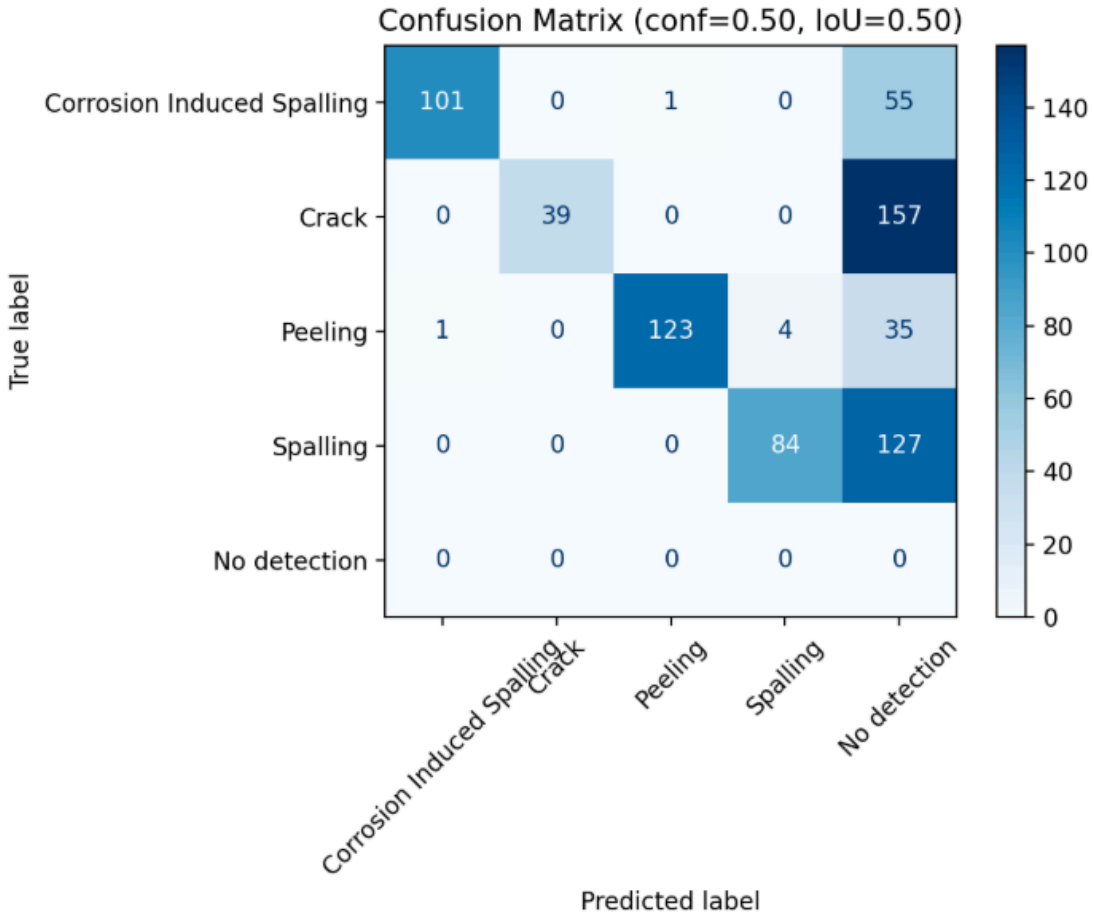
- **Crack detection is weak**, meaning engineers should not rely solely on this model for structural crack inspection.
- **Spalling detection is moderate**, suggesting improvements are possible with more data or refined labeling.
- **Overall**, the model is suitable for **broad defect screening** in civil infrastructure but may require class-specific tuning, especially for crack detection.

Key Takeaway:
The YOLOv12 segmentation model shows strong potential for structural inspection, particularly in detecting **spalling and peeling defects**, but requires further improvement for **fine crack detection** to ensure structural safety in civil engineering contexts.

Confidence 0.2 and IOU = 0.5



Confidence 0.5 and IOU = 0.5


Confusion Matrix (conf=0.50, IoU=0.50)
Corrosion Induced Spalling
Crack
Peeling
Spalling
No detection
True label
101 0 1 0 55
0 39 0 0 157
1 0 123 4 35
0 0 0 84 127
0 0 0 0 0
Corrosion Induced Spalling
Crack
Peeling
Spalling
No detection
Predicted label
140
120
100
80
60
40
20
0