DEEP LEARNING APPROACHES FOR RAILROAD INFRASTRUCTURE MONITORING:

COMPARING YOLO & VISION TRANSFORMERS FOR DEFECT DETECTION

Advay Chandramouli, Hwapyeong Song, Mingyan Liu, Aayush Damai, Husnu S. Narman, Ph.D., and Ammar Alzarrad Ph.D

MARSHALL
College of Engineering and Computer Sciences

2025 IEEE 16th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)

October 22-24, 2025

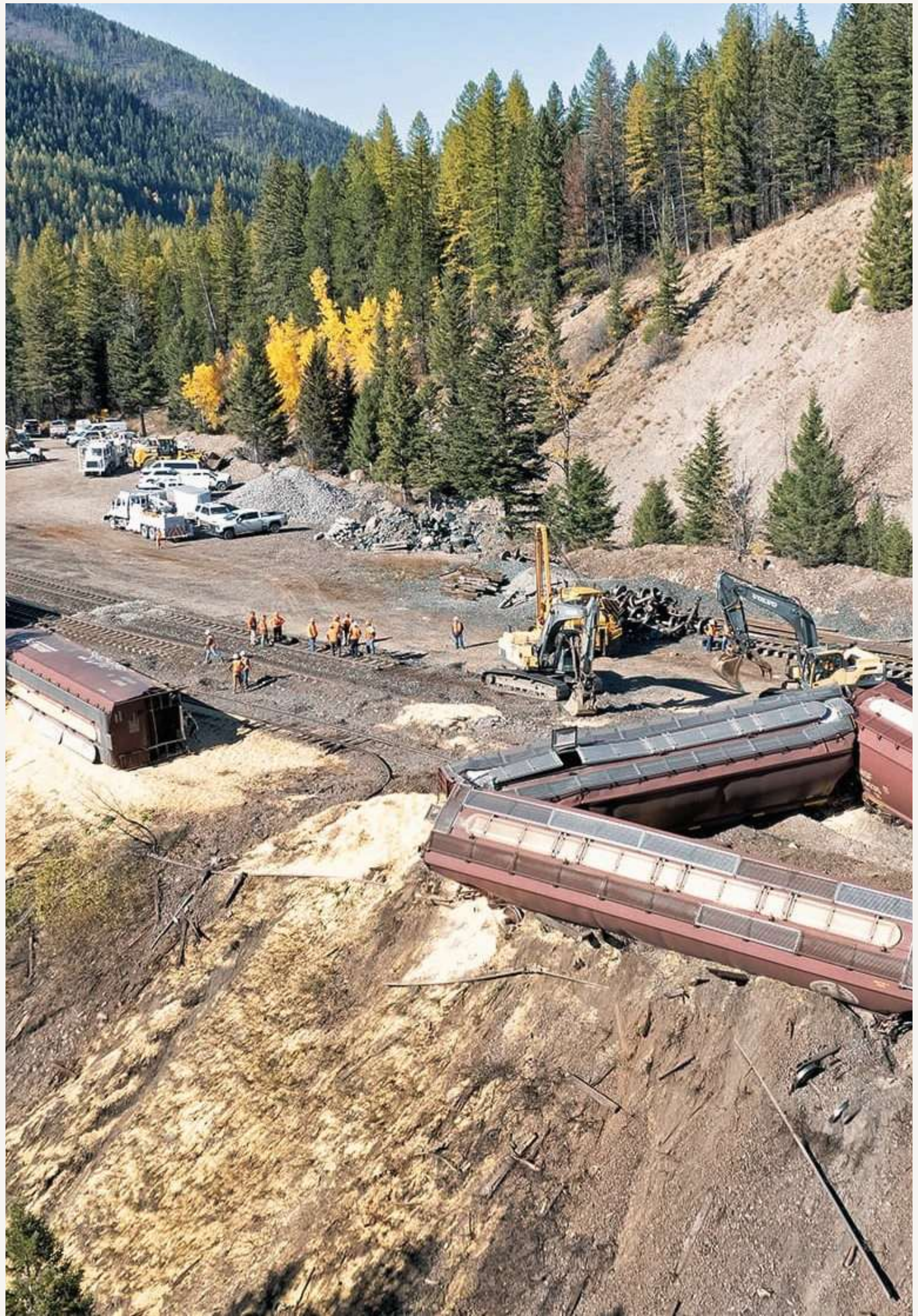
Overview

College of Engineering and Computer Sciences

Introduction

28.6+ million passengers
rely on railroads for transportation, according to Amtrak passenger data from 2023 [1]

1000+ derailments
in 2022, due to rail defects exacerbating track geometry and structural integrity [2], [3]



Current Methods & Approaches

- **Traditional inspection methods**
  like Magnetic Flux Leakage (MFL) or Ultrasonic Testing (UT) limited by speed & accuracy [4-6]

- **Early Machine Learning (ML) applications**
  using decision trees, SVMs and logistic regression models show promise in this domain for feature extraction [7], [8]

- **Latest Deep Learning (DL) advancements**
  in object detection have enabled real-time visual defect detection [9], [10]

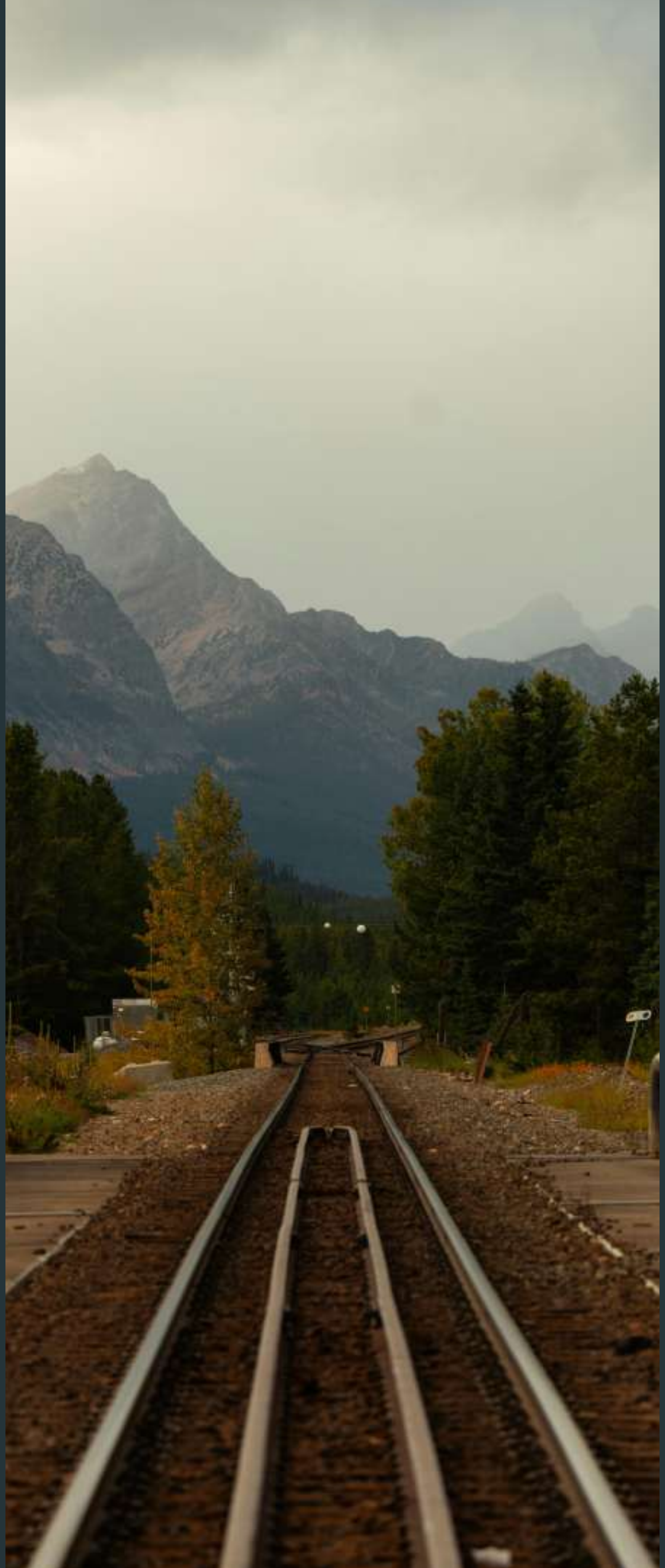
College of Engineering and Computer Sciences

Gap in Scholarly Literature

Extensive work on CNN-based object detectors
focusing on component-level detection for bolts, rails and fasteners [9], [10]

Vision Transformers (ViTs) show promise
in general object detection, but remain underexplored in railroad defect detection contexts [11], [12]

No studies directly benchmark
CNN-based object detectors and Transformer-based models in this domain



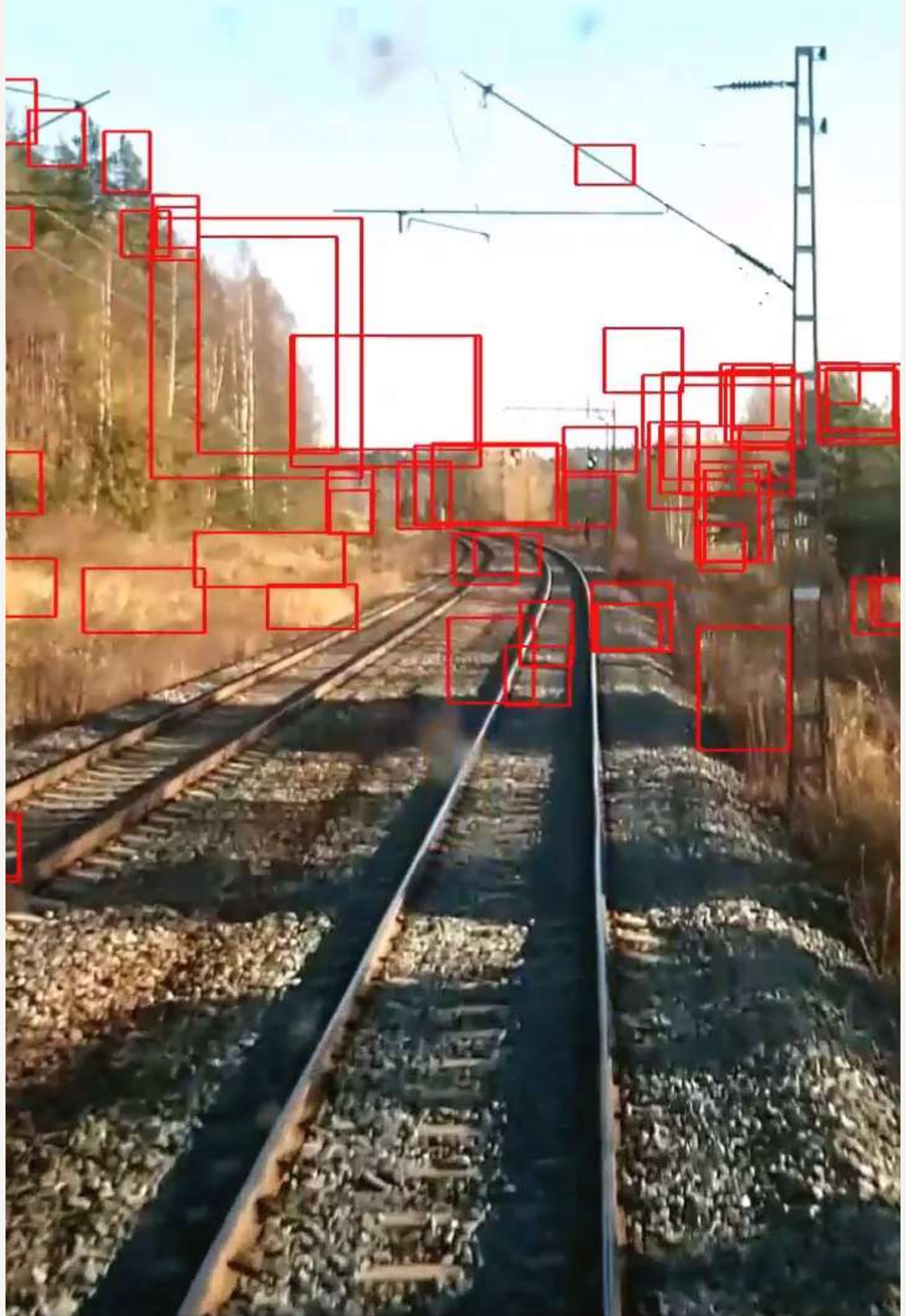
Research Question

How can deep learning-based object detection models be leveraged to detect defective railroad ties?

Machine Learning Models

Non-trivial selection, owing to the bevy of computer vision algorithms, resulting in the following study criterion:

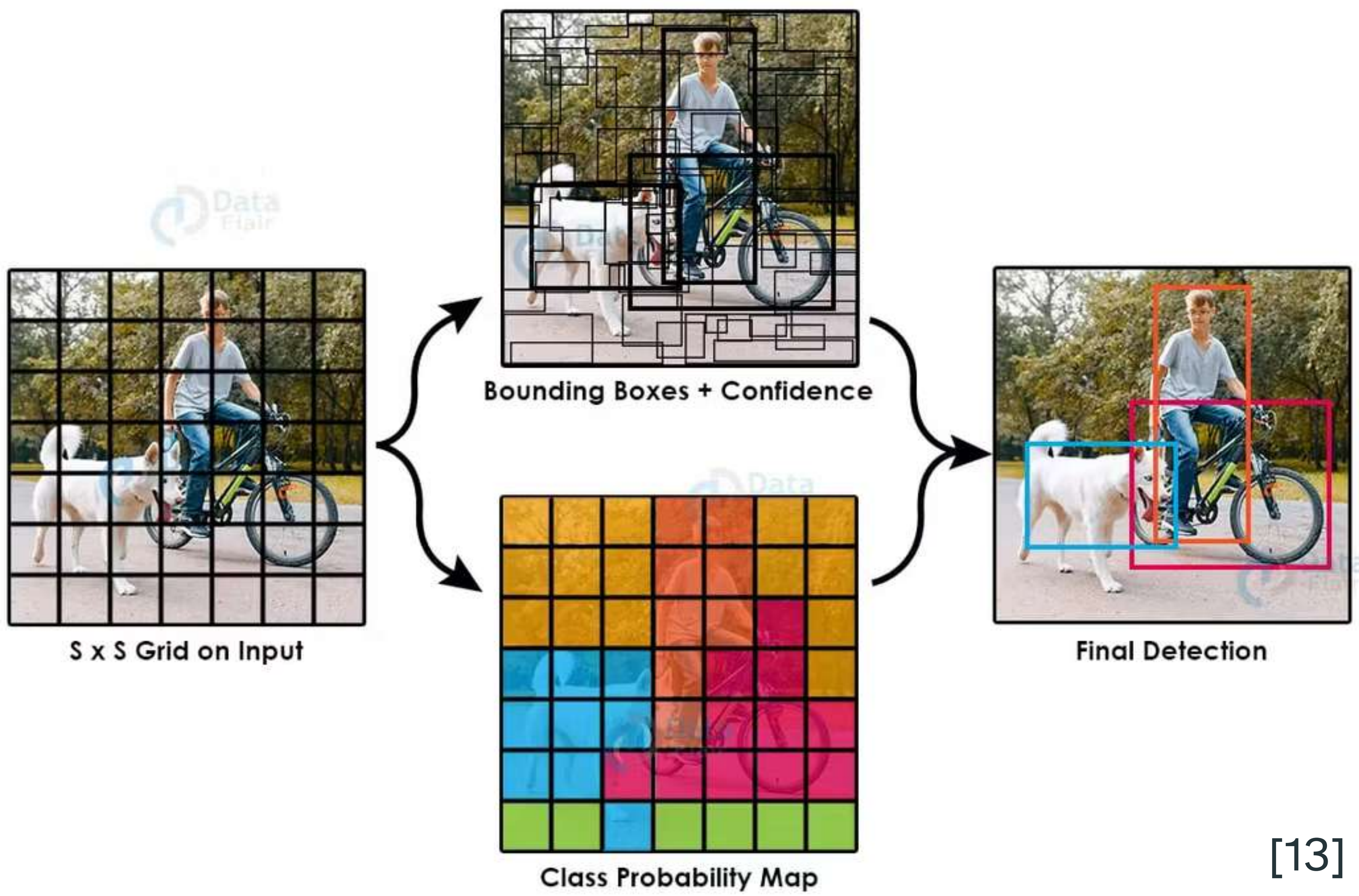
- *Supports real-time, multi-class detection tasks*
- *Effectively balances detection speed and localization accuracy*
- *Suitable for lightweight deployment environments*



MARSHALL
College of Engineering and Computer Sciences

You Only Look Once (YOLO)

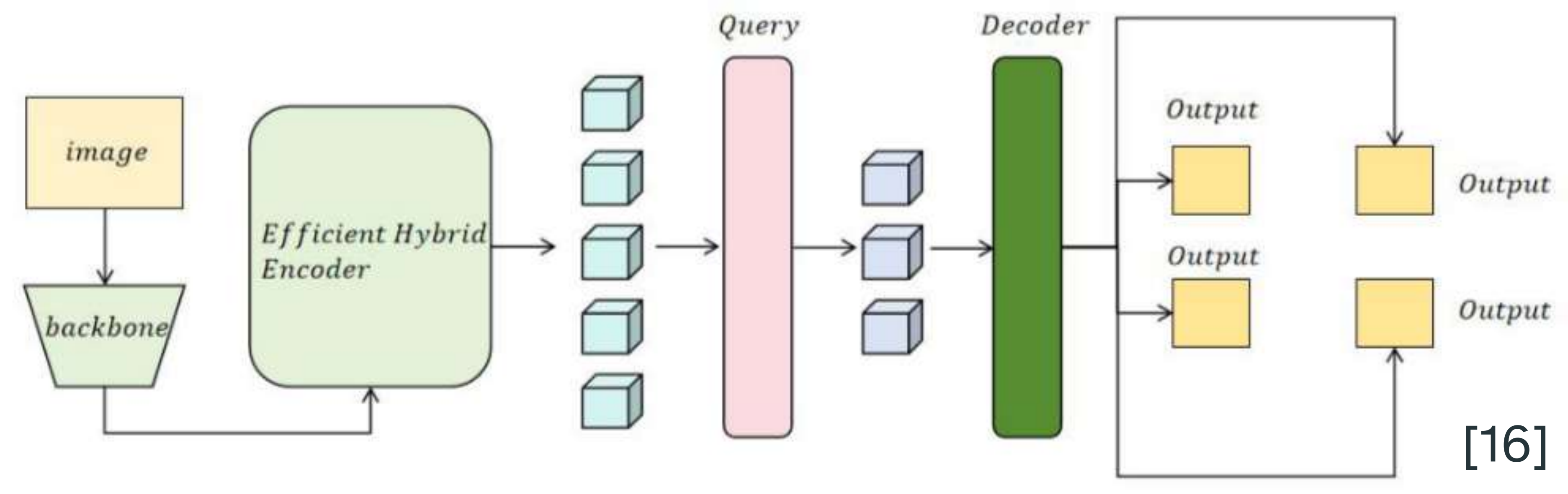
- Single-stage CNN detector
  - Partitions and processes whole image at once, enabling real-time detection [13]
- Outputs bounding boxes & confidence scores; object localization handled using Non-Maximum Suppression (NMS)
- YOLOv11 achieves higher mAP than YOLOv8 despite 22% fewer parameters [14]



[13]

Real-Time Detection Transformer (RT-DETR)

- Transformer architecture adapted for vision tasks, featuring hybrid encoder/decoder pipeline
  - Minimal NMS and learned object queries enables real-time detection [15]
- Captures both global context & fine-grained features, adapts well with limited or imbalanced data [12]



[16]

Note: *Convolutional backbone (e.g., ResNet) responsible for feature map extraction before transformer processing.*

YOLOv11 Model Comparison

Model	Params (M)	Speed - CPU (ms)	Speed - T4 GPU (ms)
YOLO11n	2.6	56.1	1.5
YOLO11s	9.4	90	2.5
YOLO11m	20.1	183.2	4.7
YOLO11l	25.3	238.6	6.2
YOLO11x	56.9	462.8	11.3

RT-DETR Model Comparison

Version	Params (M)	Speed - T4 GPU (ms)
RT-DETR-L	32.9	8.8
RT-DETR-X	~67	13.5

Note: *Parameter counts for RT-DETR vary with the chosen backbone (e.g., ResNet-50 vs ResNet-101). Values shown here are representative benchmarks.*

- Notable trade-offs between model size, speed, and accuracy:
  - Larger models deliver higher accuracy but run slower, even with GPU acceleration
  - Smaller models achieve faster inference but at the cost of accuracy
- Both YOLOv11-L and RT-DETR-L provide comparable parameter counts and satisfy this study's selection criteria

Methodology

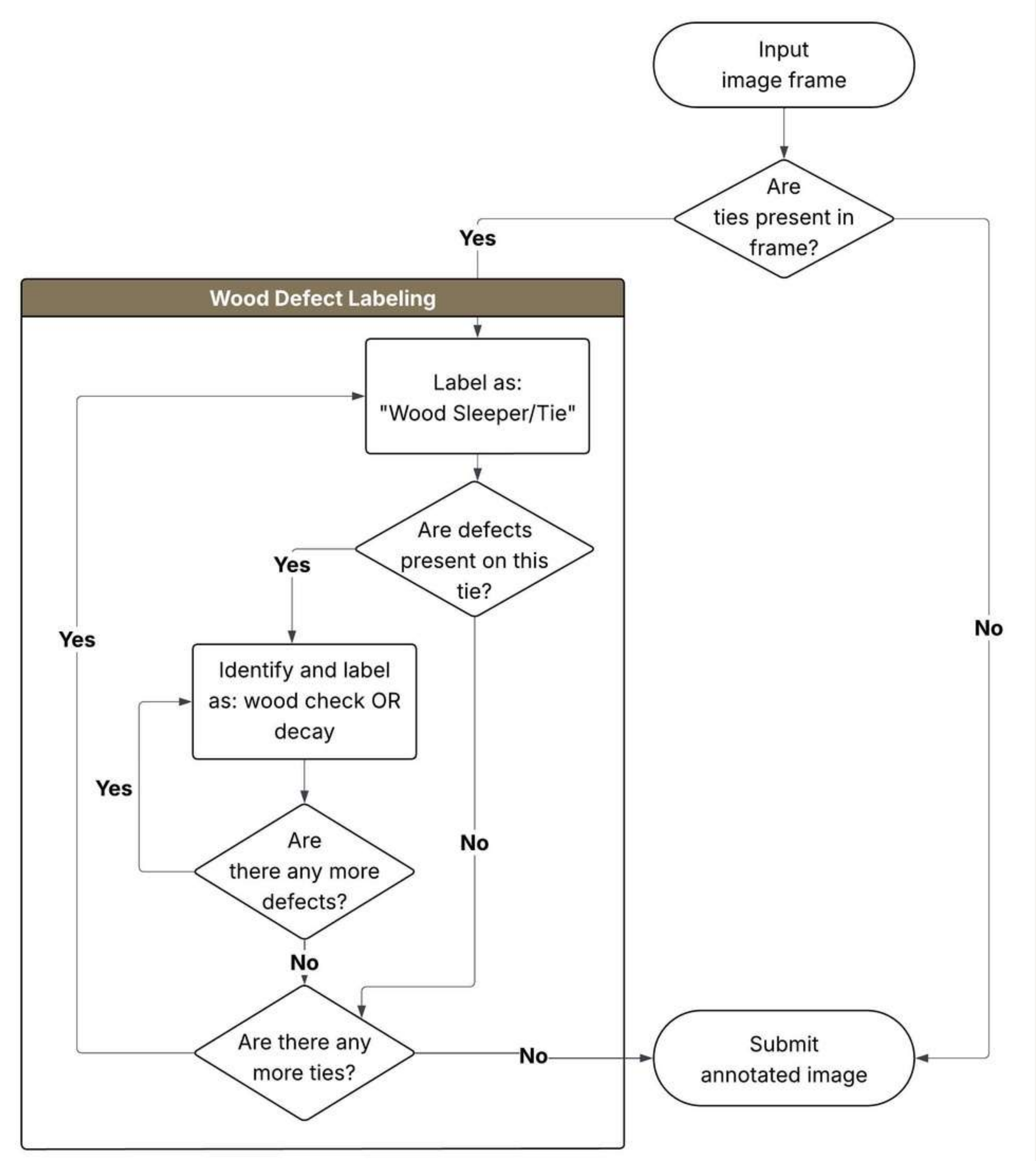
Dataset Overview:

- Collected overhead railroad footage using a custom-built camera rig
  - Extracted 573 frames from 5 minute contiguous video streams
- Sampled every 15th frame to ensure variance and a distinct set of ties per image

Dataset → Preprocessing & Annotation → Training & Evaluation Protocol

Data Preprocessing & Annotation

- Filtered dataset to <u>500 images</u> by removing blurred, over/underexposed, or obstructed frames
  - Padded to square dimensions while maintaining original 4:3 aspect ratio
- Annotated each image for following labels, using nested bounding boxes retained spatial context between ties and defects
  - *Wood Ties*
  - *Wood Checks*
  - *Wood Decay*
  - *Missing Ties*



MARSHALL
COLLEGE OF ENGINEERING AND COMPUTER SCIENCES

Sample Annotations

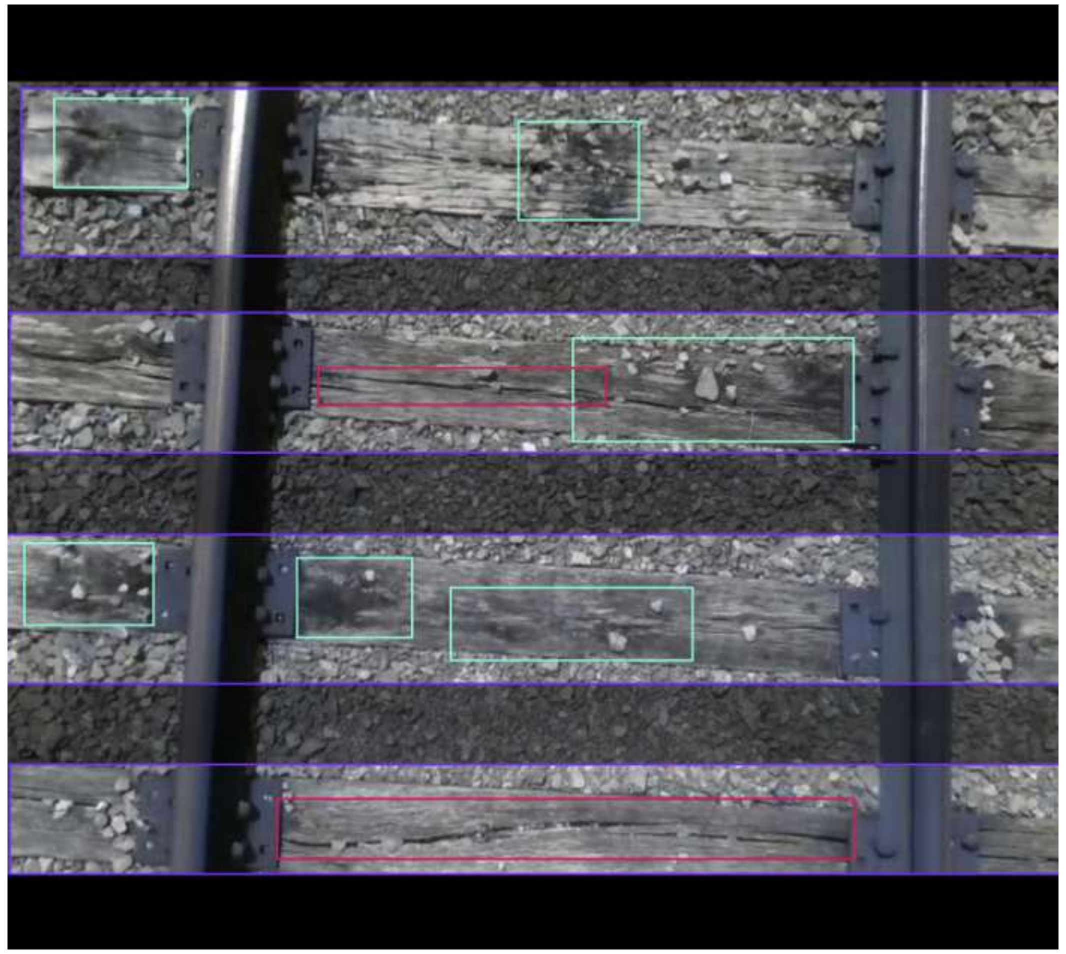
I. Missing Ties

II. Defective Ties

Wood Tie

Wood Decay

Wood Check



Data Augmentation

- Applied following image transformations to augment dataset:
    - *Contrast Stretching*
    - *Horizontal Flips*
    - *±10° hue,*
    - *±10% saturation,*
    - *±5% brightness*
    - *Salt-and-pepper noise (0.1% pixels)*

Class Label	Class Distribution
Wood Check	716
Wood Decay	1,329
Wood Ties	1,779

Training & Evaluation Protocol

I. Model Training

- **Training & Validation:** Fixed hyperparameter configuration with 5-fold cross-validation with 80/20 train–test split
- **Hardware:** NVIDIA A100 GPUs (40 GB memory, 432 tensor cores)

II. Evaluation Metrics

- F1 Score
- Precision
- Recall
- Mean Average Precision
  - IoU: 0.5 and 0.5-0.95

Results

Post-Training & Validation Evaluations:

- Plotted box loss, classification (CLS) loss, distribution focal loss over epoches for both models
    - Indicative of strong convergence, minimal overfitting and effective generalization to unseen data

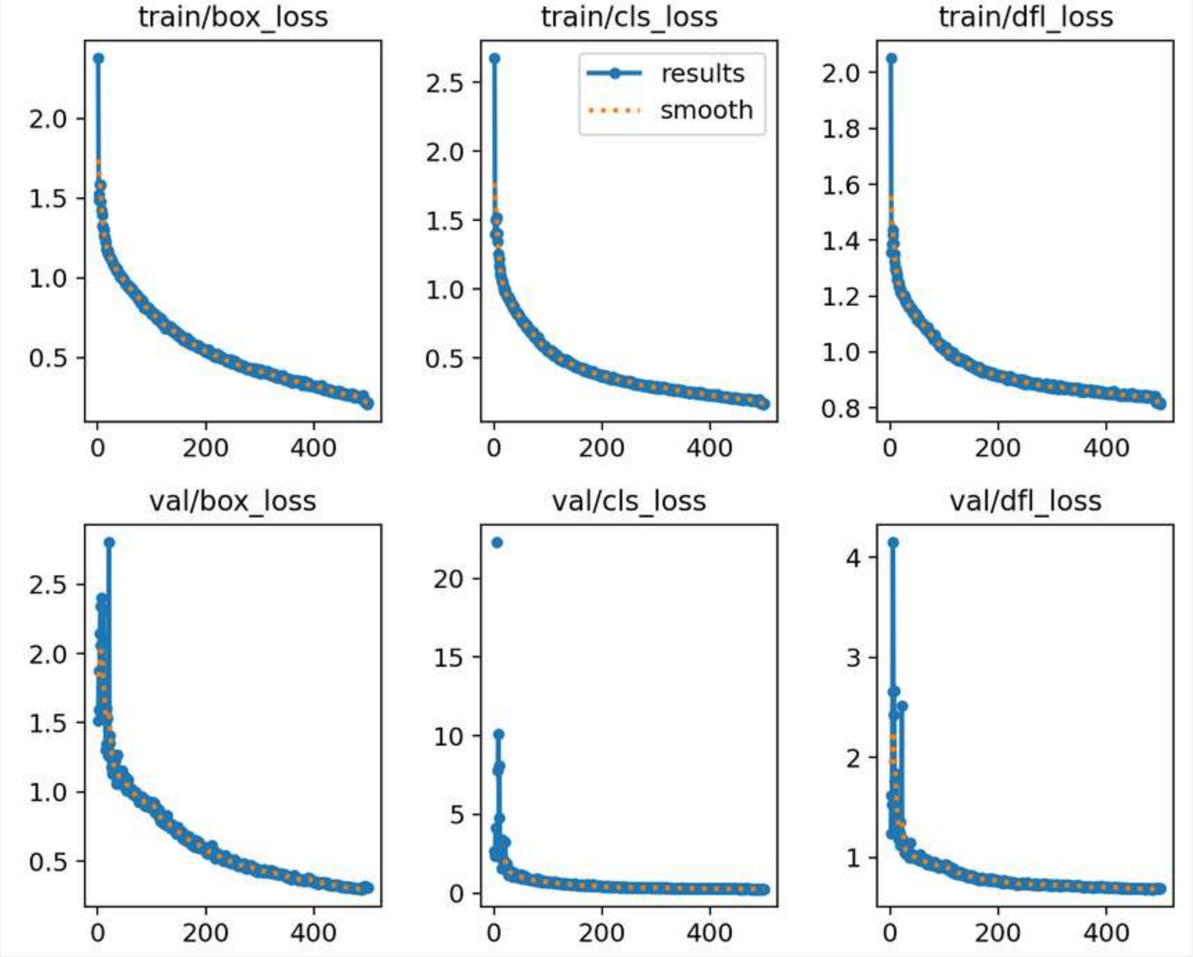


Figure A: *Loss graphs from best performing YOLOv11 Fold*

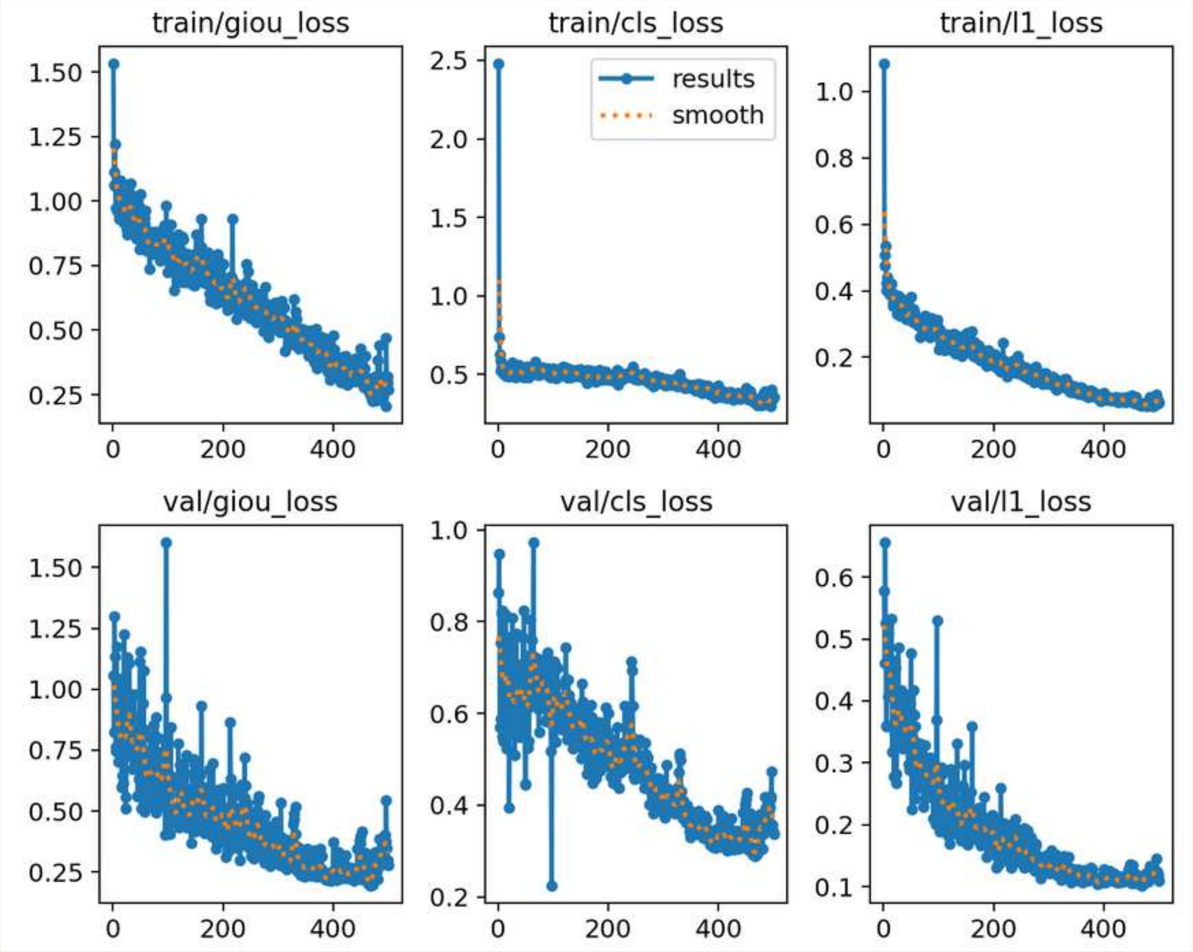


Figure B: *Loss graphs from best performing RT-DETR Fold*

Quantitative Evaluation

Metric	YOLOv11-Large	RT-DETR-Large	Winner
F1 Score	0.9400 ± 0.0089	0.9300 ± 0.0114	YOLOv11
Precision	0.9696 ± 0.0077	0.9498 ± 0.0088	YOLOv11
Recall	0.9104 ± 0.0147	0.9119 ± 0.0152	RT-DETR
mAP50	0.9530 ± 0.0106	0.9321 ± 0.0094	YOLOv11
mAP50-95	0.9014 ± 0.0134	0.7898 ± 0.0131	YOLOv11

Error Analysis

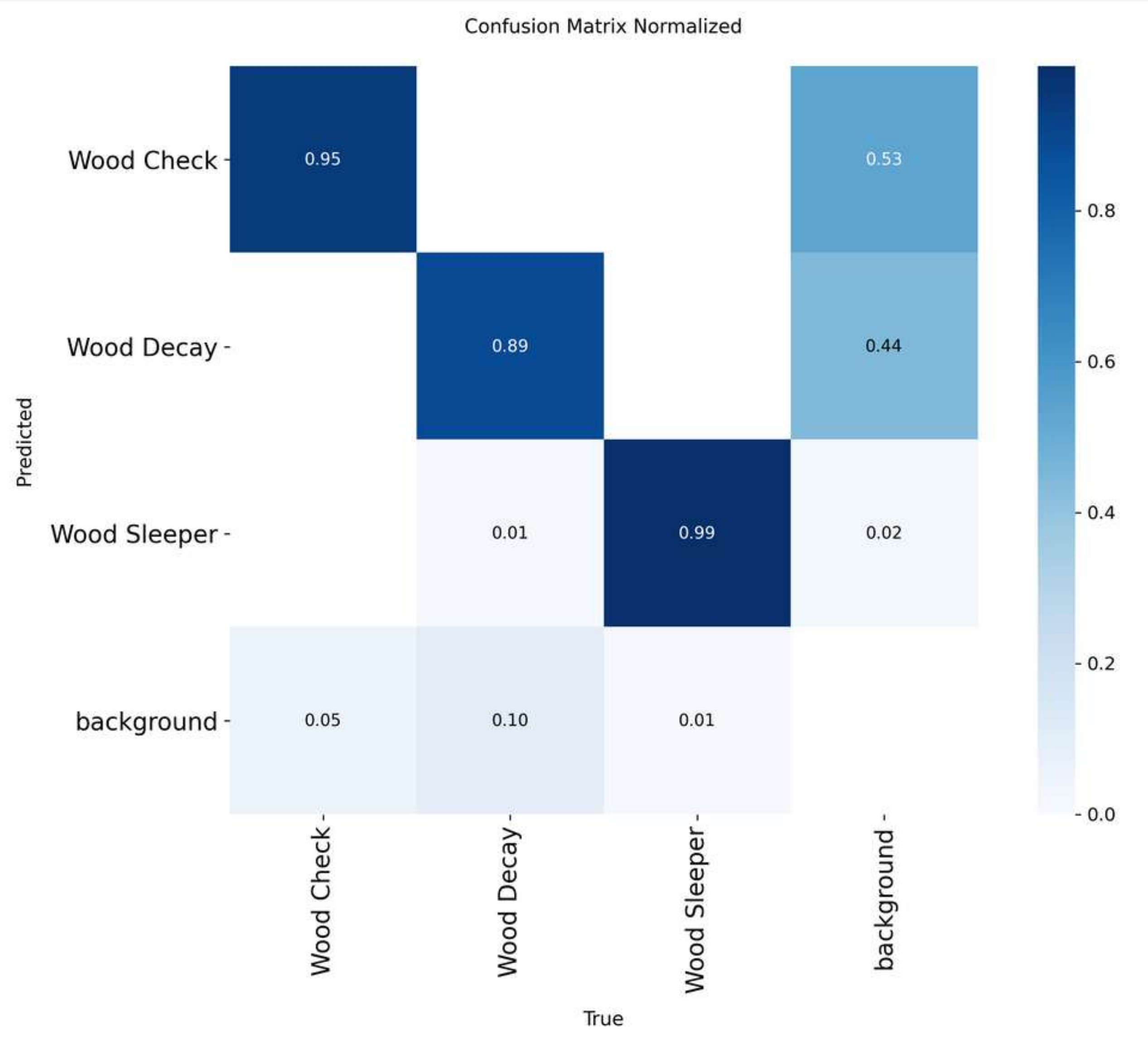


Figure C: *Normalized confusion matrix for best-performing YOLOv11 fold; more balanced false positives, fewer extreme misclassifications*

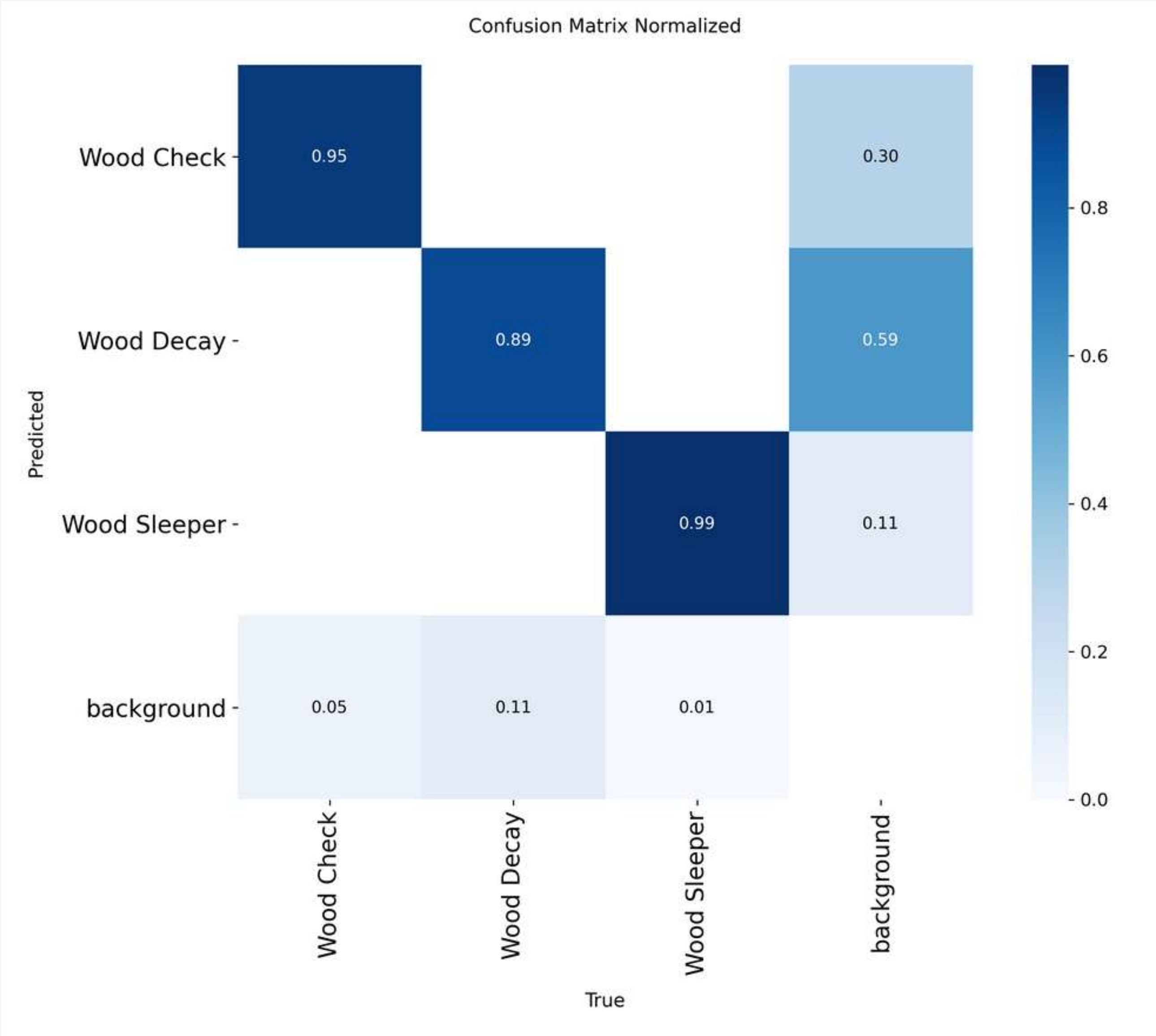


Figure D: *Normalized confusion matrix for best-performing RT-DETR fold; higher recall with inflated instances of misclassifying background as defects*

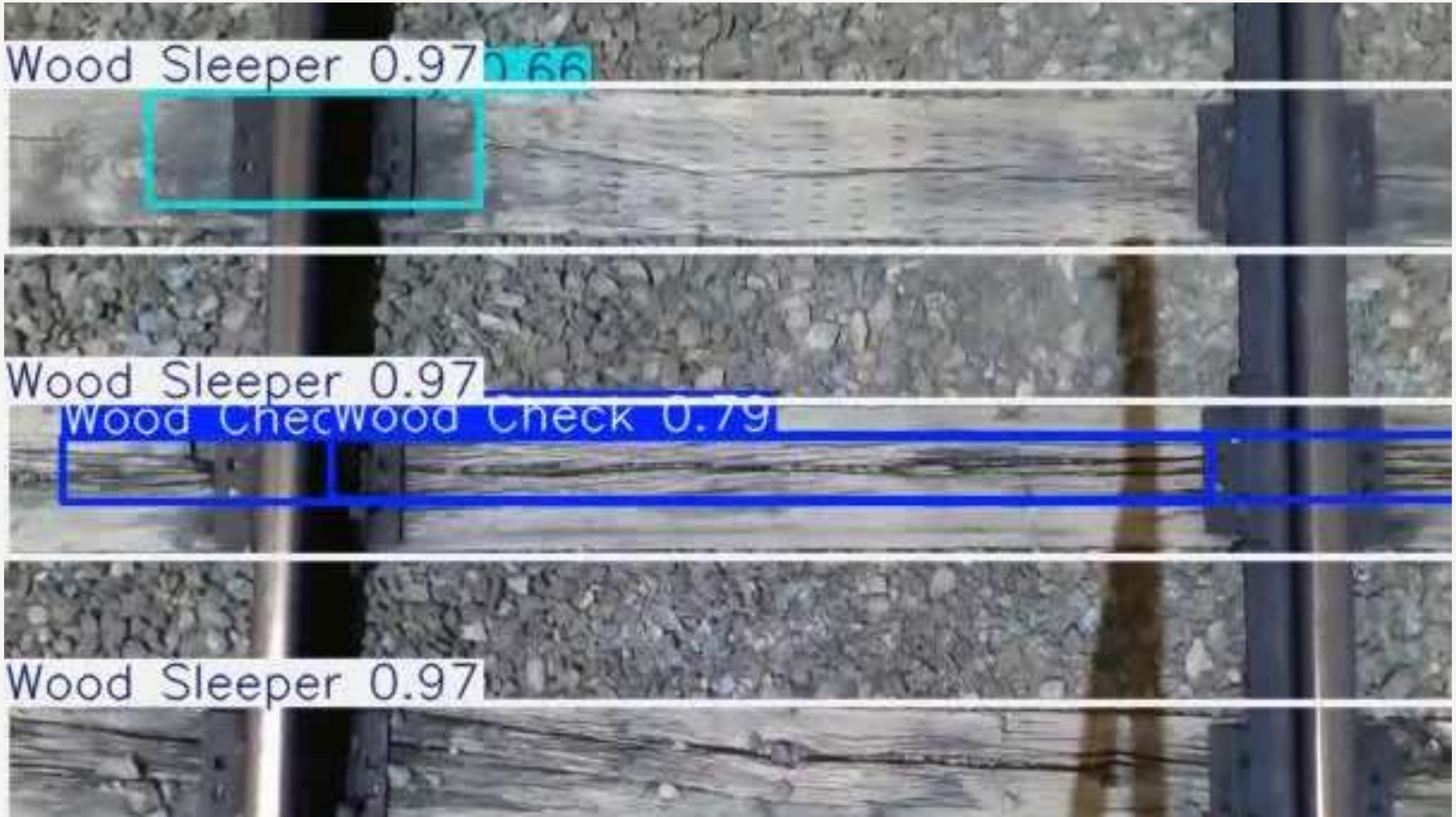
Per-Class Performance

- Per-class detection accuracy averaged across 5 folds:
  - Both models achieve strong performance on Wood Ties (~99%)
  - RT-DETR performs slightly better on *Checks*; YOLOv11 outperforms on *Decay*

Class Label	YOLOv11	RT-DETR
Wood Check	0.90	0.92
Wood Decay	0.85	0.876
Wood Ties	0.99	0.992

MARSHALL College of Engineering and Computer Sciences

Model Demonstration



COLLEGE OF ENGINEERING AND COMPUTER SCIENCES

Conclusions

- Object detection models effectively capture discrete defects and spatial relationships between tie conditions
- High-quality, consistent data paramount to maximizing DL's predictive power
- While transformer models (e.g., RT-DETR) show promise, CNN-based detectors (YOLOv11) remain superior for real-time speed, accuracy, and deployment

Limitations & Future Directions

- Address dataset constraints: field test footage captured on abandoned rail segments in West Virginia, limiting environmental variability (lighting, weather, defect types)
- Explore segmentation-based two-stage pipelines for finer localization
- Develop a defect severity rating system with governing bodies to translate model outputs into actionable maintenance insights

References

[1] Amtrak, “Amtrak Fiscal Year 2023 ridership exceeds expectations as demand for passenger rail soars,” Nov. 2023. [Online]. Available: https://media.amtrak.com/2023/11/amtrak-fiscal-year-2023-ridership-exceeds-expectations-as-demand-for-passenger-rail-soars/

[2] U.S. Department of Transportation, “Rail data dashboard,” 2023. [Online]. Available: https://data.transportation.gov/stories/s/v2un-y5se

[3] National League of Cities, “Interactive rail safety map: See derailments in communities across the U.S.,” 2023. [Online]. Available: https://www.nlc.org/resource/interactive-rail-safety-map-see-derailments-in-communities-across-the-u-s/

[4] Y. Wang, Y. Wang, P. Wang, K. Ji, J. Wang, J. Yang, and Y. Shu, “Rail magnetic flux leakage detection and data analysis based on double-track flaw detection vehicle,” Processes, vol. 11, p. 1024, 03 2023.

[5] W. Gong, M. F. Akbar, G. N. Jawad, M. F. P. Mohamed, and M. N. A. Wahab, “Nondestructive testing technologies for rail inspection: A review,” Coatings, vol. 12, p. 1790, 11 2022.

[6] Y. Xia, S. W. Han, and H. J. Kwon, “Image generation and recognition for railway surface defect detection,” Sensors, vol. 23, pp. 4793–4793, 05 2023.

[7] S. M. Mirzaei, A. Radmehr, C. Holton, and M. Ahmadian, “In-motion, non-contact detection of ties and ballasts on railroad tracks,” Applied Sciences, vol. 14, pp. 8804–8804, 09 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/19/8804

[8] J. Sresakoolchai and S. Kaewunruen, “Railway defect detection based on track geometry using supervised and unsupervised machine learning,” Structural Health Monitoring, vol. 21, p. 147592172110444, 01 2022.

[9] A. Damai, H. Song, H. S. Narman, A. Lambert, and A. Alzarrad, “Enhancing railway safety: A machine learning approach for automated detection of missing track bolts,” in Proc. ASCE Int. Conf. Computing in Civil Engineering (i3CE), New Orleans, LA, May 11–14 2025.

[10] A. D’Arms, H. Song, H. S. Narman, N. C. Yurtcu, P. Zhu, and A. Alzarrad, “Automated railway crack detection using machine learning: Analysis of deep learning approaches,” in Proc. 2024 IEEE 15th Annu. Inf. Technol., Electron. and Mobile Commun. Conf. (IEMCON), Berkeley, CA, Oct. 24–26 2024.

[11] Y. Zhao, Z. Liu, D. Yi, X. Yu, X. Sha, L. Li, H. Sun, Z. Zhan, and W. J. Li, “A review on rail defect detection systems based on wireless sensors,” Sensors, vol. 22, p. 6409, 01 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/17/6409

[12] A. Phaphuangwittayakul, N. Harnpornchai, F. Ying, and J. Zhang, “Railtrack-davit: A vision transformer-based approach for automated railway track defect detection,” Journal of Imaging, vol. 10, pp. 192–192, 08 2024.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, real-time object detection,” arXiv preprint arXiv:1506.02640, 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1506.02640

[14] M. L. Ali and Z. Zhang, “The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection,” Computers, vol. 13, no. 12, 2024. [Online]. Available: https://www.mdpi.com/2073-431X/13/12/336

[15] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, “Detrs beat yolos on real-time object detection,” 2024. [Online]. Available: https://arxiv.org/abs/2304.08069

[16] W. He, Y. Zhang, T. Xu, Y. Liang, and T. An, “Object Detection for Medical Image Analysis: Insights from the RT-DETR Model,” arXiv preprint arXiv:2501.16469, Jan. 2025.

COLLEGE OF ENGINEERING AND COMPUTER SCIENCES

Thank You

To the National Science Foundation's Research Experiences for Undergraduates (REU) Program, Marshall University, and the U.S. Army Engineer Research & Development Center (ERDC) for supporting this work.

Q&As