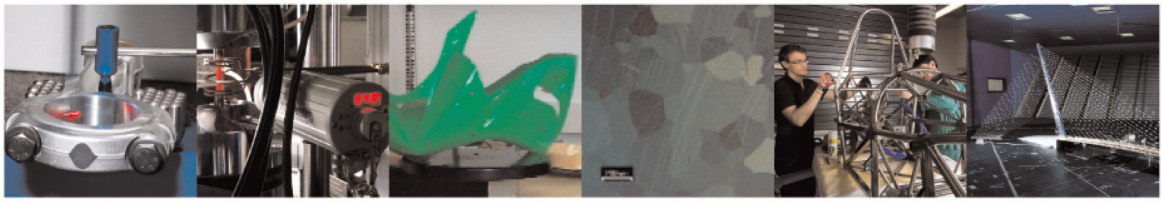
POLITECNICO
MILANO 1863

DIPARTIMENTO DI MECCANICA

mecc



Enhancing railway infrastructure monitoring using hybrid VTOL drones: a case study on inspection and surveillance using custom YOLOv12 object detector

Cardellicchio, Angelo;Faccini, Leonardo;Nitti, Massimiliano;Renò, Vito;Tarsitano, Davide;Zappa, Emanuele

This is a post-peer-review, pre-copyedit version of Angelo Cardellicchio, Leonardo Faccini, Massimiliano Nitti, Vito Renò, Davide Tarsitano, and Emanuele Zappa "Enhancing railway infrastructure monitoring using hybrid VTOL drones: a case study on inspection and surveillance using custom YOLOv12 object detector", Proc. SPIE 13570, Multimodal Sensing and Artificial Intelligence for Sustainable Future, 135701C (4 August 2025); https://doi.org/10.1117/12.3062378

Enhancing Railway Infrastructure Monitoring using Hybrid VTOL Drones: a case study on Inspection and Surveillance using custom YOLOv12 Object Detector

Angelo Cardellicchio[a], Leonardo Faccini[b], Massimiliano Nitti[a], Vito Renò[a], Davide Tarsitano[b], and Emanuele Zappa[b]

[a]Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing (STIIMA) National Research Council of Italy, Bari, Italy
[b]Department of Mechanical Engineering, Politecnico di Milano, Milano, Italy

ABSTRACT

Hybrid VTOL drones represent a promising solution for patrolling railway lines for security purposes and conducting macroscopic infrastructure integrity checks. These systems can be an alternative to multi-rotor drones, especially in applications requiring less detailed analysis, such as mapping, macro-object identification, or obstacle detection. To achieve this purpose, a model based on a specifically tailored revision of YOLO12 was trained on a previously gathered dataset, acquired to model different use cases in real-world applications, and tested, laying the basis for an end-to-end embedded framework able to provide railway segmentation and obstacle 3D reconstruction via an end-to-end processing pipeline based on deep learning and photogrammetry. The results demonstrated the feasibility of the proposed methodology under real-case scenarios and laid the basis to be combined with segmentation methods and photogrammetry to provide a complete processing pipeline for the automatic, real-time identification of relevant obstacles on railway tracks.

1. INTRODUCTION

In the last decade, the use of unmanned aerial vehicles (UAVs), or drones, for the condition monitoring of civil infrastructures has widely increased, and railway lines are no exception.[1] In this context, UAVs have been effectively used to monitor specific infrastructure assets such as bridges,[2] switches and signaling systems,[3] catenaries,[4] etc., and for obstacle detection.[5] Some attempts have been also made to perform detailed analyses as the identification of rail surface defects.[6]

Most of these applications rely multi-rotor drones due to they ease of maneuverability and relatively low cost. Additionally, their ability to hover enables the capture detailed images of the specific target from different points of view.[7] However, their limited flight time and speed make them unsuitable for the continuous monitoring on railway lines, which requires operation along extended routes and greater distances.

Fixed-wing UAVs appear to be better suited for this type of continuous monitoring, thanks to their higher flight speed and longer operational range. However, because they operate like conventional airplanes and must remain above their stall speed,[7] their use in railway applications is generally limited to large-scale mapping of the infrastructure.[8]

Hybrid Vertical Take-Off and Landing (VTOL) drones offer a promising solution for railway monitoring, as they combine the extended range, speed, and payload capacity of fixed-wing UAVs with the vertical take-off and landing capability of multi-rotor systems. These hybrid drones can serve as alternatives to multi-rotor UAVs for various monitoring tasks, especially in scenarios requiring relatively long flight capabilities to quickly and autonomously reach the target point of the line, and in cases where high-resolution inspection is not essential—such as mapping, identifying large objects, or detecting obstacles.

Authors are listed alphabetically and all authors equally contributed to this work.
Further author information: (Send correspondence to Vito Renò)
Vito Renò: E-mail: vito.reno@stiima.cnr.it

This paper aims to explore new possibilities in inspection, surveillance, and maintenance through the use of VTOL drones operated in Beyond Visual Line of Sight (BVLOS) mode. The study focuses on two main objectives: patrolling railway lines for security and conducting macroscopic infrastructure integrity checks to ensure safe train circulation.

Nowadays there is a high demand of real-time object detection tasks that play a crucial role in a high number of practical applications because of their applicability at high TRL values, implying that AI models must achieve an optimal trade-off between inference speed and detection accuracy. To perform this tasks, the YOLO (You Only Look Once) models in all their iterations have been succesfully used in a large number of application domains, demonstrating state-of-the-art performance and high context adaptability of the approach. Real time evaluation of vision systems using YOLO models (from v6 to v11) in terms of their efficiency on various platforms, including the employment of optimization techniques, has been recently performed for an unmanned surface vehicle videos and sensor analysis.[9] Even if on one hand YOLO models have been constantly improved during the time (leading to multiple major model versions) for example including attention mechanisms in the model architecture that shown some inefficiencies in their traditional formulation, limiting their integration into high-speed detection frameworks. The latest version of YOLO – YOLOv12[10,11] – tries to overcome these limitations by incorporating advanced attention-based strategies within the YOLO architecture, enhancing both computational efficiency and detection performance.

The rest of the paper is organized as follows: in Section 2 details about the studied dataset as well as YOLOv12 architecture and implementation details are reported; Section 3 reports the experiments and the results achieved while Section 4 concludes the paper and describes possible future works.

2. MATERIALS AND METHODS

2.1 Dataset

To achieve the monitoring goal and lay the basis for an on-board, real-time processing system, a dataset composed of several videos acquired under different settings was first gathered in testing railway areas. More specifically, the videos were recorded using commercial cameras at different illuminations conditions and viewpoints, providing both daylight and night time video frames. A total number of $N = 50613$ frames were then extracted from the videos and used to feed a zero-shot object detection method, to provide an initial labelling assessment for the following classes: Grass, Pallet, Railway, Wood.

These labels were then refined by domain experts and made the knowledge base used in the experiments for the characterization and evaluation of a tailored version of YOLOv12 object detector model. Examples of labelled frames are reported in Figure 1.

2.2 YOLOv12 model

This paragraph resumes the main innovation of YOLOv12 model for object detection, that was used in the experiments to recognize the abovementioned categories of objects. Compared to previous versions of the model, YOLOv12 introduces an architecture particularly centered on the attention mechanism, moving beyond traditional CNN-based approaches but with the aim of keeping a real-time inference speed.[12] In particular, this version of YOLO achieves higher accuracy if compared to the previous ones, balancing speed and precision. Therefore, the main YOLOv12 innovations are:

- **Area Attention Mechanism**, a novel self-attention approach able to process large receptive fields reducing computational complexity, based on a equally-sized non-overlapping region-based approach. More specifically, with respect to the classic self attention, the complexity is reduced by computing the area attention on small portions of the feature map – via a straightforward reshape operation – that is partitioned into L segments of size $(\frac{H}{L}, W)$ or $(H, \frac{W}{L})$, overcoming the requirement of window partitioning methods seen in other attention models such as Shifted Window,[13] Criss-Cross Attention,[14] or Axial Attention[15]

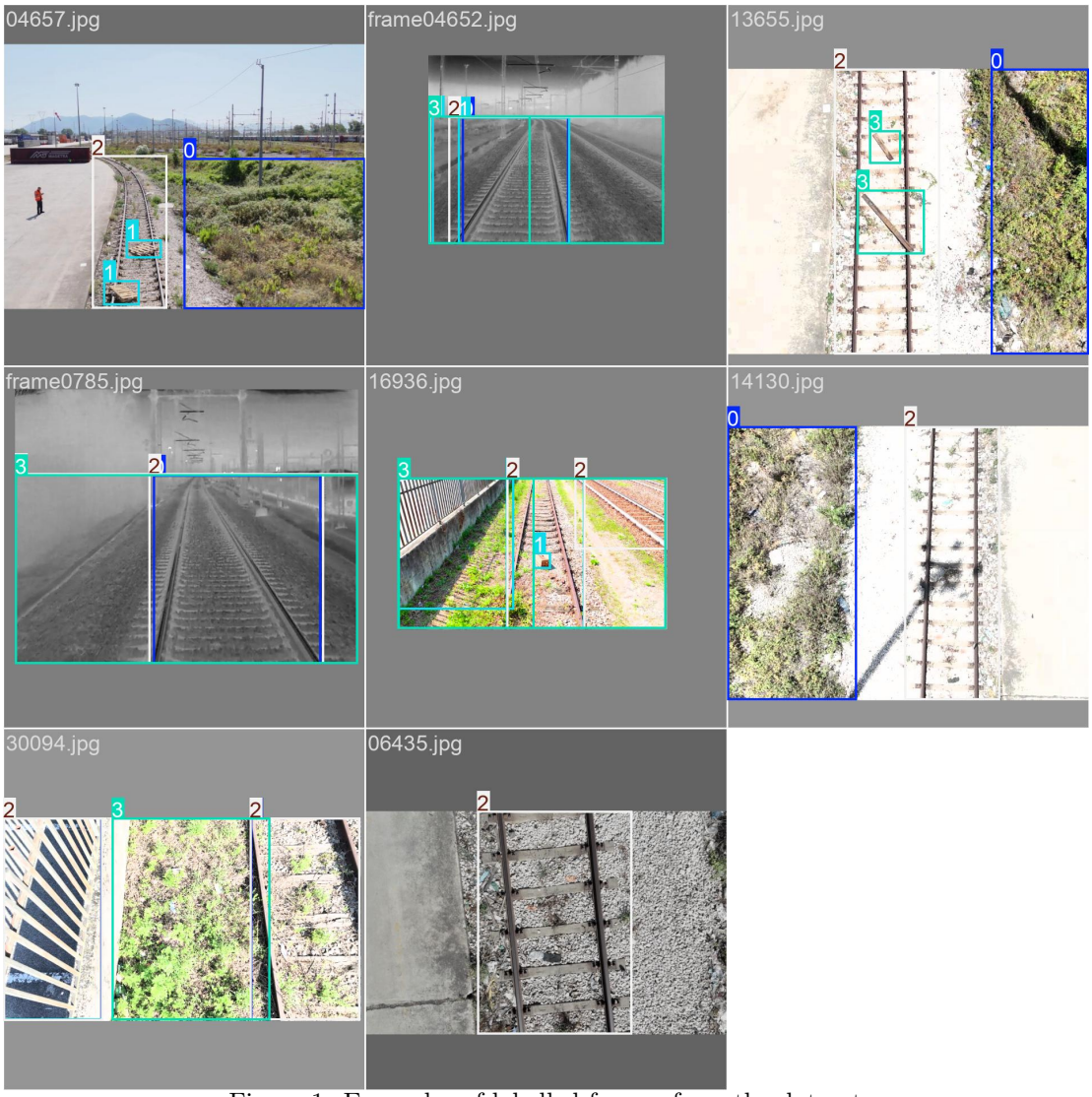


Figure 1: Examples of labelled frames from the dataset

- **Residual Efficient Layer Aggregation Networks (R-ELAN)**, that is used to perform the aggregation of the features optimizing the performance in models that leverage attention mechanisms, incorporating block-level residual connections inspired by Vision Transformers[16] and a bottleneck-like structure. More specifically, in previous YOLO versions ELAN was implemented[17] but introduced drawbacks regarding gradient blocking and optimization difficulties (more marked in deeper models), R-ELAN is different because it improves stability and convergence using residual connections to generate a comprehensive feature map before feeding it to the bottleneck layers, enhancing the overall efficiency with the redundancies avoidance.[11]

Among other additional improvements and refinements specifically tailored to boost the model performance it is worth highlighting the removal of explicit positional encoding that is achieved by using 7×7 separable convolutions in the attention module (called position perceiver) as well as a GPU optimized implementation of the attention called FlashAttention that make the overall complexity of the attention more similar to convolutional neural networks in terms of speed.[12]

2.3 Evaluation metrics

The following subsection reports the evaluation metrics for the results of the experiments computed in this work. In particular, several metrics commonly used in object detection were employed, in particular *precision* (P), *recall* (R), $F1 - score$ ($F1$).

2.3.1 Positives and Negatives

Before extending the evaluation to a multi-class problem, it is essential to consider a scenario where only two types of objects need to be recognized: a binary classification problem. In this case, it is possible to define a *positive* and *negative* class, leading to the introduction of the following quantities:

- **TP** represents the *true positives*, the instances correctly identified by the model as belonging to the positive class.
- **TN** represents the *true negatives*, the instances correctly identified by the model as belonging to the negative class.
- **FP** represents the *false positives*, that is, the instances of the negative class incorrectly identified by the model as belonging to the positive class.
- **FN** represents the *false negatives*, that is, the instances of the positive class incorrectly identified by the model as belonging to the negative class.

In the specific case of YOLO object detectors, each detection (i.e. the bounding box BB_i) for a specific class is compared to the ground-truth $GT(BB_i)$ in terms of Intersection over Union (IoU), that is defined as:

$$IoU = \frac{BB_i \cap GT_(BB_i)}{BB_i \cup GT_(BB_i)}$$

where $GT(\cdot)$ is a function that evaluates the ground truth bounding box (if any) nearest to the generic i-th YOLO detection BB_i. $IoU \to 0$ for no-matching bounding boxes, whilst $IoU \to 1$ in case of correct matches. It is worth noting that, to compute the positive or the negative quantity described before it is essential to define a threshold based on the IoU score, that is commonly set to $\tau = 0.5$, meaning that a detection is considered good with at lease 50% overlap. For this reason, in the specific case of object detection, the performance metrics can be categorized in this way:

- **TP** refers to a correct prediction of the object detector with respect to the ground truth.
- **FP** refers to a prediction that has no matches with the ground truth $IoU \to 0$.
- **FN** refers to ground truth boxes without any prediction from the object detector.

In addition, in YOLO object detection problems these metrics are enriched considering possible background quantities, defined in the following way:

- **B-FP**, that represents the *background false positives*, i.e., the instances identified by the model on the background. Note that in this case, the model output could be a correct instance (not labelled in the original dataset) or an incorrect one.
- **B-FN**, that represents the *background false negatives*, i.e., the instances labelled in the original dataset but not identified by the model.

2.3.2 Precision, Recall, and F1 Score

Given these quantities, precision P and recall R can be computed as follows.

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

The formulation for the $F1$ score can be directly derived from Equations 1 and 2:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

3. EXPERIMENTS AND RESULTS

In this section, the performance of different YOLOv12 models size – namely: Nano (**N**), Small (**S**), Medium (**M**) and Large (**L**) - are reported. Different model sizes correspond to different number of layers, parameters and gradients according to the following Table 1.

Table 1: YOLOv12 models recap

Model size	Layers	Parameters	Gradients
N	272	2.602.288	2.602.272
S	272	9.284.096	9.284,080
M	292	20.199.168	20.199.152
L	488	26.450.784	26.450.768

The experiments were performed using a workstation machine instance equipped with an AMD EPYC 9654 Processor, 128 GB of RAM, and two NVIDIA H100 GPU with 96 GB of RAM each. All the source code was implemented exploiting the Ultralytics library with Python 3.11.

The dataset has been split using a 70/30 proportion in training and test set. All the models have been fine-tuned starting from the pretrained version available in the Ultralytics repository. Figure 2 shows the performance of the models computed on the test set. First of all, it is worth observing that the object detection performance of YOLOv12 is directy proportional to the model size/complexity, meaning that bigger models show better results. All the upper left 4×4 submatrices extracted from the confusion matrices are diagonally dominant matrices, thus indicating the discriminative power of the YOLOv12 object detector in our application scenario that is coherent independently from the specific model size. However, for all the models, there is a non neglectable presence of B-FPs and B-FNs instances of Grass, Pallet, Railway or Wood labelled bounding boxes that suggest the need of increasing data variability in the dataset. Looking at the F1-score values with respect to the confidence parameter of YOLO, there is no significant improvement of the confidence threshold even when the model is the most complex one. The F1 value is $\sim 94\%$ for all the models. For the specific use case considered in this work, due to the limited set of objects to be recognized and to the high similarity of the frames included in the dataset, a less complex model can be effectively considered for the deployment as the performance gain – either in terms of correct class detection increase and B-FP/B-FN reduction – is limited.

4. CONCLUSION AND FUTURE WORKS

In this paper a case study about the inspection and surveillance of railway infrastructure using custom deep-learning object detector models based on YOLOv12 architecture has been presented. The scope of the work was focused on the characterization of such AI model architectures, with different sizes and trainable parameters, on a custom dataset acquired at different illumination settings and viewpoints, namely during the day and the night, from a hybrid VTOL drone. The findings confirmed the applicability of the evaluated architectures in

real-world situations, opening the door to integrating it with segmentation and photogrammetry techniques, creating a comprehensive system for a potential real-time automated detection of railway track hazards. Future improvements of this work will be devoted to the extension of the dataset and experiments on a higher number of railway tracks and routes, as well as the rationalization, optimization and implementation of the AI model on dedicated hardware for a better scalability and deployment.

ACKNOWLEDGEMENTS

This study was carried out within the MOST – Sustainable Mobility National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the views and opinions of the authors. Neither the European Union nor the European Commission can be considered responsible for them. The authors would like to thank Mr. Michele Attolico and Dr. Paola Romano for their technical support.

REFERENCES

[1] Flammini, F., Pragliola, C., and Smarra, G., "Railway infrastructure monitoring by drones," in [*2016 International Conference on Electrical Systems for Aircraft, Railway, Ship Propulsion and Road Vehicles amp; International Transportation Electrification Conference (ESARS-ITEC)*], IEEE (Nov. 2016).

[2] Narazaki, Y., Hoskere, V., Chowdhary, G., and Spencer, B. F., "Vision-based navigation planning for autonomous post-earthquake inspection of reinforced concrete railway viaducts using unmanned aerial vehicles," *Automation in Construction* **137**, 104214 (May 2022).

[3] Mittal, S. and Rao, D., "Vision based railway track monitoring using deep learning," *arXiv preprint arXiv:1711.06423* (2018).

[4] Geng, Y., Pan, F., Jia, L., Wang, Z., Qin, Y., Tong, L., and Li, S., "Uav-lidar-based measuring framework for height and stagger of high-speed railway contact wire," *IEEE Transactions on Intelligent Transportation Systems* **23**, 7587–7600 (July 2022).

[5] Guan, L., Li, X., Yang, H., and Jia, L., "A visual saliency based railway intrusion detection method by uav remote sensing image," in [*2020 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*], 291–295, IEEE (Aug. 2020).

[6] Bojarczak, P. and Lesiak, P., "Uavs in rail damage image diagnostics supported by deep-learning networks," *Open Engineering* **11**, 339–348 (Jan. 2021).

[7] González-Jorge, H., Martínez-Sánchez, J., Bueno, M., Arias, and Pedor, "Unmanned aerial systems for civil applications: A review," *Drones* **1**, 2 (July 2017).

[8] Jarrett, C., Perry, K., and Stol, K., "Controller comparisons for autonomous railway following with a fixed-wing uav," in [*2015 6th International Conference on Automation, Robotics and Applications (ICARA)*], 104–109, IEEE (Feb. 2015).

[9] Mela, J. L. and Sánchez, C. G., "Yolo-based power-efficient object detection on edge devices for usvs," *Journal of Real-Time Image Processing* **22**, 108 (May 2025).

[10] Tian, Y., Ye, Q., and Doermann, D., "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524* (2025).

[11] Tian, Y., Ye, Q., and Doermann, D., "Yolov12: Attention-centric real-time object detectors," (2025).

[12] Khanam, R. and Hussain, M., "A review of yolov12: Attention-based enhancements vs. previous versions," *arXiv preprint arXiv:2504.11995* (2025).

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], 10012–10022 (October 2021).

[14] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W., "Ccnet: Criss-cross attention for semantic segmentation," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], (October 2019).

[15] Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., and Guo, B., "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 12124–12134 (June 2022).

[16] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and Jégou, H., "Going deeper with image transformers," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], 32–42 (October 2021).

[17] Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M., "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 7464–7475 (June 2023).

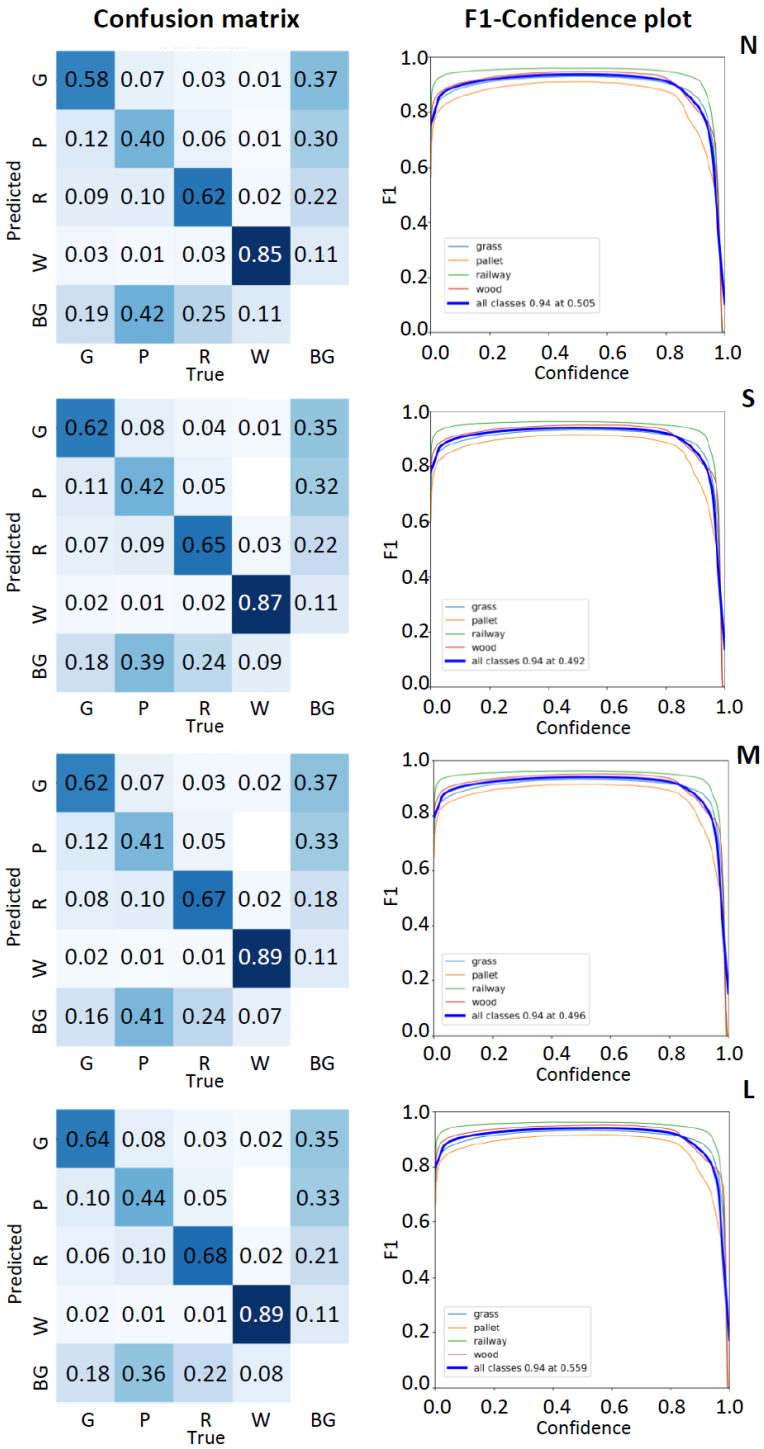


Figure 2: Confusion matrix and F1 vs. Confidence score plot. For the confusion matrices, BG refers to the Background class, G to Grass, P to Pallet, R to Railway and W to Wood. YOLOv12 model size is reported in the upper right part of each subfigure