European Research Studies Journal
Volume XXVIII, Issue 4, 2025

pp. 1044-1057

The Impact of YOLOv8 and YOLOv9 Neural Network Architectures on the Quality and Speed of Object Detection in Assistive VR/MR Applications for Visually Impaired Users

Submitted 19/09/25, 1st revision 14/10/25, 2nd revision 26/10/25, accepted 20/11/25

Piotr Lichograj[1]

Abstract:

Purpose: *This paper examines the performance and applicability of lightweight deep-learning architectures for real-time object detection in mixed-reality (MR) and virtual-reality (VR) environments designed to assist blind and visually impaired individuals. The primary objective is to determine which model—YOLOv8 or the lightweight YOLOv9-t variant—provides the superior balance of accuracy, sensitivity, and processing speed, particularly for on-device deployment on wearable devices such as the Meta Quest 3 headset. This research addresses the growing need for efficient, edge-based AI solutions that enhance navigation and spatial awareness without relying on cloud computing, thereby ensuring user privacy and low latency.*
Design/Methodology/Approach: *To conduct the study, a custom dataset comprising 15,000 annotated images was curated from a 30-minute video captured from a first-person perspective in a public building setting. This dataset includes 17 object classes critical for indoor navigation, such as persons, doors (open, closed, glass), stairs (up and down), elevators, furniture items (chairs, desks, benches, sofas), and amenities (reception, cafeteria, cloakroom, trash bins, vending machines, computers). The images reflect realistic challenges, such as varying lighting conditions, occlusions, and a predominance of small or distant objects, which mimic everyday scenarios faced by visually impaired users. The dataset was split into training and testing subsets to ensure comparability, with augmentations applied, including random scaling, flipping, and brightness adjustments to improve model robustness. Both YOLOv8 and YOLOv9-t were trained under identical conditions, utilising the same hyperparameters, image resolution (640 × 640), and augmentation strategies. Evaluation metrics encompassed precision, recall, F1-score, mean Average Precision at IoU 0.5 (mAP@0.5), and across a range of thresholds (mAP@[0.5:0.95]), alongside detailed F1-confidence curve analysis. Post-training, the models were converted to the ONNX format for cross-platform compatibility and deployed in a Unity Passthrough environment on the Meta Quest 3, leveraging Vulkan and NNAPI for hardware-accelerated inference. This setup enabled real-time testing in an MR context, where detected objects are projected into the 3D scene with bounding boxes, facilitating immediate multimodal feedback.*
Findings: *Results indicate that YOLOv9-t attains comparable overall accuracy to YOLOv8, with mAP@0.5 scores of 0.965 versus 0.969, respectively. However, YOLOv9-t achieves its peak F1-score at a lower confidence threshold (approximately 0.40 compared to 0.45 for YOLOv8), leading to enhanced recall and fewer missed detections without a substantial increase in false positives. This behaviour is particularly advantageous in assistive MR systems, where prioritising safety—such as alerting users to potential hazards like stairs or*

[1] *John Paul II University in Biala Podlaska, Poland, ORCID: 0000-0002-1351-4874, e-mail: p.lichograj@gmail.com;*

obstacles—is more critical than eliminating minor false alarms. Additionally, YOLOv9-t demonstrates faster training convergence (stabilising after 60 epochs versus 100 for YOLOv8) and reduced inference latency on mobile hardware, making it more efficient for battery-constrained wearables. The findings position YOLOv9-t as the optimal choice for MR-based assistive navigation tools. Its robustness at lower thresholds supports advanced features, such as graded multimodal cues (e.g., initial haptic vibrations escalating to audio descriptions). At the same time, on-device processing aligns with ethical standards for data privacy. By integrating technical evaluations with usability and ethical considerations, this study contributes to the advancement of inclusive AI technologies that promote greater independence, mobility, and safety for visually impaired individuals.
Practical implications: *As one of the initial empirical comparisons of YOLOv8 and YOLOv9-t in authentic MR settings, this work emphasises human-centred design principles. It offers practical insights for developing AI-driven applications in fields like education (e.g., virtual training simulations), rehabilitation (e.g., mobility therapy), and healthcare (e.g., daily living aids).*
Originality/Value: *Future research could explore ensemble methods or integrate depth-sensing for even more precise spatial mapping, ultimately fostering more innovative and more accessible environments in smart buildings and public spaces.*

Keywords: *Assistive artificial intelligence; visual impairment support; human–machine interaction; rehabilitation technology; health informatics.*

JEL codes: *I15, O33, L86.*

Paper type: *Research article.*

1. Introduction

VR/MR-based navigation systems represent a promising approach for assistive technologies that aid blind and visually impaired users. The primary function of such systems is to semantically "explain" the environment in real time: detect and label objects, estimate their position, and deliver information to the user through sound or haptic cues.

Algorithmically, this requires low latency, moderate computational demand, and high recognition quality under challenging conditions—namely, variable lighting, occlusions, and a predominance of small objects. The YOLO family (Redmon et al., 2016) established the standard for single-pass detection, combining classification and localisation into a single prediction, thereby reducing latency to levels suitable for real-time interfaces. Subsequent generations—from YOLOv4/v5 through YOLOv8 to YOLOv9—have improved both architecture (backbone, head) and training engineering (schedules, augmentations, losses), with growing emphasis on lightweight and edge-AI models (Bochkovskiy *et al.,* 2020; Jocher *et al.,* 2024; Wang *et al.,* 2024). For blind users, recall is critical: missing stairs, a doorway frame, or a service vehicle in a corridor is more dangerous than a few false alarms.

A functional MR system does not necessarily need to maximise mAP in every class; instead, it must balance precision and recall to minimise "silence" in real-world risk situations. Hence, beyond comparing mAP and F1, analysing F1–F1-confidence curves and model behaviour at low confidence thresholds is essential. This paper focuses on YOLOv8 and YOLOv9-t—two detectors with similar overall quality but distinct operational characteristics. YOLOv9-t performs stably at lower thresholds and learns faster, facilitating maintenance and frequent retraining in production environments.

2. Literature Review

In the domain of object detection, models are traditionally categorised into two-stage and single-stage architectures. Two-stage detectors, such as Faster R-CNN (Ren *et al.,* 2017), generate region proposals before classification, often achieving higher accuracy but at the cost of increased computational complexity and latency. In contrast, single-stage detectors such as SSD (Liu *et al.,* 2016) and the YOLO family (Redmon *et al.,* 2016) perform bounding box regression and classification in a single pass, prioritising real-time performance suitable for edge devices.

This trade-off between speed and accuracy has been extensively analysed, with studies showing that modern convolutional detectors must balance these factors for practical applications (Huang *et al.,* 2017). Datasets like Microsoft COCO (Lin *et al.,* 2014) have become benchmarks for evaluating these models, emphasising challenges such as small objects and occlusions prevalent in real-world scenarios.

The YOLO series has undergone significant evolution to address these challenges. Starting with YOLOv4 (Bochkovskiy *et al.,* 2020), improvements in bag-of-freebies (e.g., data augmentation) and bag-of-specials (e.g., activation functions) enhanced speed and accuracy. YOLOv8 introduced an anchor-free head and improved feature transfer between resolution levels, enhancing small-object detection and reducing anchor-related complexity (Jocher *et al.,* 2024).

Building on this, YOLOv9 introduced Programmable Gradient Information (PGI) and the Generalised Efficient Layer Aggregation Network (GELAN), which improves parameter efficiency and stability—especially in the "tiny" and "small" variants (Wang *et al.,* 2024). Recent variants, such as YOLOv10, further optimise for real-time vision on edge devices, demonstrating superior performance in assistive contexts (Wang *et al.,* 2024).

In the context of assistive technologies for visually impaired users, VR/MR systems integrate object detection with multimodal feedback to enhance navigation and spatial awareness. Early systems focused on sensor-based aids (e.g., wearable devices for obstacle avoidance; Tapu *et al.,* 2020), but recent works have leveraged deep learning for a more semantic understanding.

For instance, DeepNAVI (Kim *et al.,* 2022) utilises smartphone-based detection to provide audio cues for obstacles, thereby improving mobility. Similar approaches include YOLOv8-based systems for real-time object recognition in smart glasses and assistive sticks, often deployed on Raspberry Pi or Android platforms for on-device inference (Park *et al.,* 2025).

A systematic review highlights that XR (extended reality) technologies, including MR, can simulate environments for training visually impaired users, reducing real-world risks (Fogli *et al.,* 2025). Multimodal systems that combine detection, depth estimation, and haptic/audio feedback have been shown to enhance user safety and agency, with YOLO variants such as YOLOv7 and YOLOv8 excelling in obstacle detection for navigation assistance (Shah *et al.,* 2023; Maheshwari *et al.,* 2024). Challenges include low-light performance, which is addressed by enhanced YOLO models in complex indoor/outdoor settings (Rahman *et al.,* 2019).

Ethical considerations are crucial in these applications. Frameworks like UNESCO's Recommendation on AI Ethics (2023) emphasise proportionality, privacy, and no-harm principles, ensuring on-device processing to protect user data. IEEE's Ethically Aligned Design (2023) advocates for human-centred AI, prioritising accessibility and transparency in assistive technologies. Reviews on AI in VR for blind users stress the need for inclusive design to avoid overstimulation (Lee *et al.,* 2025). Future directions include adaptive thresholds and ensembles for better recall in safety-critical scenarios.

3. Methodology

3.1 Dataset and Data Preparation

The dataset was collected in a public building using a first-person perspective video lasting over 30 minutes. Every 5th frame was sampled, producing 15,000 images with varied lighting, occlusions, and object distances. Manual annotation covered 17 classes: person, open door, closed door, glass door passage, reception, stairs up, stairs down, elevator, cafeteria, cloakroom, chair, desk, bench, trash bin, vending machine, sofa/couch, computer.

The dataset is unbalanced—people and furniture dominate, while “cafeteria” and “vending machine” appear less frequently. This imbalance and the predominance of small objects reflect the realistic environment in which the system operates.

Data were split into training and test sets using a fixed split for comparability. Images were resized to 640×640 with standard augmentations, including random scaling, horizontal flipping, brightness/contrast changes, and mild colour distortions—to improve robustness to indoor variations.

3.2 Architectures and Training

Both YOLOv8 (used as the qualitative reference model) and YOLOv9-t (the lightweight variant of YOLOv9) were evaluated under identical experimental conditions. The same input configuration was applied in both cases, including image resolution, augmentation techniques, and a comparable training regimen. YOLOv8 was trained for 100 epochs, whereas YOLOv9-t reached metric stabilisation after 60 epochs, confirming its faster convergence.

The impact of the Programmable Gradient Information (PGI) and Generalised Efficient Layer Aggregation Network (GELAN) mechanisms is evident in the reduced occurrence of informational bottlenecks and improved learning efficiency on challenging examples, such as tiny objects and low-illumination regions within the frame.

Both architectures were evaluated using a common set of standard object detection metrics: precision [1], Recall [2], F1-score [3], mAP@0.5, and mAP@[0.5:0.95], as well as precision–recall (PR) curves, F1–F1-confidence curves, and confusion matrices. These indicators collectively provide a comprehensive view of model effectiveness.

Unlike traditional classification tasks, where a decision pertains to a single image, object detection requires evaluating both positional accuracy and semantic correctness—that is, the degree of overlap between predicted and ground-truth bounding boxes, along with the proper class assignment.

$$\text{"Precision"}=TP/(TP+FP) \quad [1]$$

$$\text{"Recall"}=TP/(TP+FN) \quad [2]$$

$$F1=2\times(\text{"Precision"} \times\text{"Recall"})/(\text{"Precision"} +\text{"Recall"}) \quad [3]$$

Where:

- TP (True Positives) – correctly detected objects with matching class and IoU above threshold.
- FP (False Positives) – incorrect detections (nonexistent or misclassified objects).
- FN (False Negatives) – missed real objects.

These metrics form the basis for AP (Average Precision), the area under the Precision–Recall curve for a class. The mean across all classes is the mean Average Precision (mAP). IoU (Intersection over Union) [4] measures overlap between prediction and ground truth; detections above the IoU threshold are counted as true positives.

$$IoU=(|B\_pred \cap B\_gt|)/(|B\_pred \cup B\_gt|) \quad [4]$$

In practical object detection, IoU thresholds of 0.5 (mAP@0.5) and the average over multiple thresholds (0.5 → 0.95, step 0.05), i.e. mAP@[0.5:0.95], are used to provide a more rigorous evaluation of the model's matching accuracy.

To determine operational thresholds for MR systems, F1 maxima versus confidence were analysed. In practice, MR benefits from configurations that minimise misses while tolerating moderate "soft" detections, filtered by interface logic.

3.3 Deployment Platform and Inference Pipeline

The target deployment environment was based on integration with Unity Passthrough, which enables the real-time projection of camera imagery into a 3D scene, and implementation on the Meta Quest 3 headset powered by the Snapdragon XR2 Gen 2 SoC. The trained models were exported to the ONNX format and executed using ONNX Runtime Mobile, which utilises Vulkan and NNAPI acceleration.

At the core of the processing pipeline lies a continuous loop encompassing frame acquisition, neural inference, projection of bounding boxes and labels into the mixed-reality space, and generation of audio or haptic feedback. This architecture enables seamless synchronisation between visual detection and multimodal user interaction.

The use of the ONNX format facilitates further optimisation, including mobile-oriented execution profiles, operator reduction, and maintaining a single binary compatible across multiple hardware configurations.

A privacy-first design philosophy was adopted: all inference is performed on-device, ensuring that no image data leaves the headset. Event logging and analytical data are fully anonymised and remain under the user's control. This approach aligns with ethical AI design principles, enhancing both data security and user trust in assistive mixed reality systems.

4. Results and Interpretation

In the overall evaluation, the YOLOv8 model achieved an mAP@0.5 of 0.969 and an F1-score of 0.93, while the YOLOv9-t model obtained 0.965 and 0.92, respectively. The differences between these values are minor and statistically insignificant across most object classes. This indicates that both architectures exhibit comparable global detection accuracy, with the primary distinction lying in their behaviour under varying confidence thresholds.

Figure 1. *Real-time object detection in an MR environment using Passthrough mode on Meta Quest 3. The algorithm identifies navigation-relevant objects (e.g., doors, chairs) and projects bounding boxes directly onto the live camera feed in the 3D scene.*

A key aspect of the comparative analysis is the shape of the F1–Confidence curve, which illustrates how the F1 Score evolves as the detection confidence threshold increases. For YOLOv8, the optimal F1 value occurs at a confidence threshold of approximately 0.45, whereas for YOLOv9-t, the optimum appears at a lower threshold—around 0.40.

This means that YOLOv9-t maintains robust and stable performance even at reduced confidence levels. In practical Mixed Reality (MR) applications, this property translates directly into a higher recall rate and a lower risk of “silence” — that is, failing to detect or report an object that is visually present but only weakly visible (for example, distant, partially occluded, or in shadow).

When analysing the results per class, YOLOv8 demonstrates a slight advantage in categories such as person, chair, desk, elevator, and closed door. In contrast, YOLOv9-t achieves higher mAP scores in classes of greater navigational relevance for visually impaired users—such as descending stairs, canteen, cloakroom, and sofa/couch—and provides more reliable differentiation between open and closed doors. This distinction is critical for spatial orientation and safety in indoor navigation systems.

The confusion matrices for both models are largely diagonal, indicating a high overall classification accuracy. However, YOLOv9-t produces fewer confusions between semantically similar classes, which in turn reduces inconsistencies in the feedback delivered to the user through audio or haptic interfaces. From the perspective of assistive mixed reality systems, this reduction in ambiguity improves both user trust and situational awareness, as the system's responses become more consistent and contextually coherent.

Table 1. *Per-class mAP@0.5 values for YOLOv8 and YOLOv9-t models. The results are based on 17 object classes relevant to indoor navigation in public buildings. The Δ column shows the difference between YOLOv8 and YOLOv9-t for each class.*

Object Class (Polish / English)	YOLOv8	YOLOv9-t	Δ (v8 − v9-t)
Osoba / Person	0.916	0.902	+0.014
Drzwi otwarte / Open door	0.976	0.973	+0.003
Drzwi zamknięte / Closed door	0.974	0.967	+0.007
Przejście – drzwi szklane / Glass door passage	0.957	0.946	+0.011
Recepcja / Reception desk	0.984	0.985	–0.001
Schody w górę / Stairs up	0.972	0.969	+0.003
Schody w dół / Stairs down	0.973	0.977	–0.004
Winda / Elevator	0.992	0.985	+0.007
Stołówka / Canteen	0.982	0.989	–0.007
Szatnia / Cloakroom	0.989	0.993	–0.004
Krzesło / Chair	0.963	0.951	+0.012
Biurko / Desk	0.986	0.978	+0.008
Ławka / Bench	0.970	0.964	+0.006
Kosz na śmieci / Trash bin	0.967	0.958	+0.009
Automat z przekąskami / Vending machine	0.991	0.983	+0.008
Sofa / Kanapa / Sofa / Couch	0.969	0.973	–0.004
Komputer / Computer	0.917	0.908	+0.009

Source: *Author's calculations.*

The differences between YOLOv8 and YOLOv9-t are marginal, fluctuating within approximately ±0.01 in mAP@0.5 values. YOLOv8 maintains a slight advantage in classes characterised by a large number of training samples and simpler visual structures—such as person, desk, and chair. In contrast, YOLOv9-t yields higher performance in categories that are critical for navigational safety, including descending stairs, canteens, cloakrooms, and sofas.

It is worth noting that despite YOLOv9-t achieving a marginally lower mAP overall, the model exhibits a higher recall at a lower confidence threshold (approximately 0.40). This indicates that YOLOv9-t responds more effectively under uncertain or visually ambiguous conditions—a feature of particular importance for visually impaired users relying on assistive Mixed Reality (MR) systems. In practical terms, this characteristic—greater sensitivity with minimal loss of precision—makes YOLOv9-t the more advantageous choice for real-time VR/MR applications that support spatial orientation and navigation assistance.

From an operational standpoint, YOLOv9-t also demonstrates shorter inference times and greater stability when the confidence threshold is reduced. In mobile and embedded environments, this has dual significance: first, it facilitates smooth rendering and consistent frame rates (typically 45–60 fps depending on the scene), and second, it enables the use of a more tolerant initial threshold, where detection filtering can occur at the interface level rather than within the model itself. This approach mitigates the risk of missing critical detections that could impact user safety.

In MR systems designed for the visually impaired, this operational mode aligns with a human-centred communication paradigm: the system may "whisper" soft cues about potentially relevant objects, while the multimodal interaction layer—integrating auditory and haptic feedback—decides whether and when to escalate the signal intensity (for instance, from a brief tactile pulse to a full spoken description). Such a graded response model enhances situational awareness while minimising cognitive overload, ensuring that users receive timely yet non-intrusive guidance within complex indoor environments.

4.1 Relevance for Blind Users

In assistive systems designed for visually impaired users, safety is the paramount objective — specifically, minimising instances of "silence", where the system fails to report obstacles or spatial reference points. A difference of 0.004 in mAP between models carries far less practical significance than the threshold behaviour and the shape of the F1–F1-Confidence curve.

The YOLOv9-t model, which operates stably at a confidence threshold of approximately 0.40, increases the number of detections near the visibility limit and enables the implementation of a multi-stage feedback sequence: an initial low-intensity warning signal (haptic feedback), followed after 200–300 ms by a confirmation cue, and finally, a full verbal message if the object remains in view.

This scheme reduces cognitive load while simultaneously lowering the likelihood that the system will remain silent when it should issue an alert—such as in the presence of descending stairs or closed doors. Particularly critical are semantically similar pairs of classes.

The distinction between "open door" and "closed door" determines the navigation path and whether the interface communicates "passage accessible" or "seek an alternative route".

A lower number of such semantic confusions in YOLOv9-t enhances user trust, which in turn fosters a greater willingness to delegate spatial attention to the system. This relationship aligns with findings from prior studies on mixed and augmented reality (MR/AR) and multimodal feedback in blind and visually impaired (BVI) populations (Kim *et al.,* 2022)

4.2 Interface, Multimodality and Information Design

Implementing object detection represents only the first step—equally important is how the information is communicated to the user. Mixed Reality (MR) environments enable the integration of haptic feedback, spatial audio, and minimal voice overlays, forming a multimodal feedback system. For visually impaired individuals with heightened sensory sensitivity, overstimulation poses a significant risk.

An adaptive, two-stage notification scheme, based on both detection persistence and class priority (e.g., stairs, doors, elevators > furniture), enables the reconciliation of high recall with user comfort. Such a structure ensures that critical elements of the environment are always reported, while non-essential objects generate only subtle or aggregated cues.

The user interface should also incorporate information on direction and distance, supported by depth estimation or SLAM-based spatial mapping, so that users can interpret the spatial relationship of detected objects—such as “in front,” “to the left,” or “two meters away.”

Operating effectively at a lower confidence threshold, YOLOv9-t acts as a more sensitive detector, capturing low-visibility objects without overwhelming the user. The decision to escalate feedback intensity—whether to deliver a short haptic pulse or a full verbal message—remains the responsibility of the interaction layer, ensuring both responsiveness and contextual relevance.

4.3 Ethics, Accessibility and Responsibility

Assistive technologies must adhere to the principle of “accessibility-by-design”, which implies designing systems from the outset with the needs of users with disabilities in mind, rather than attempting to “retrofit” accessibility features in the final stages of development.

In practice, this includes modular control of audio output, the ability to mute notifications quickly, personalisation of haptic intensity, silent or discreet operating modes for social settings, and explicit communication of uncertainty levels within the system’s responses.

From a privacy standpoint, on-device inference is strongly preferred: image data remains entirely within the headset, and all event logs are anonymised and fully controlled by the user. The YOLOv9-t model facilitates this architecture due to its lightweight structure and easy exportability to the ONNX format, enabling efficient local deployment without reliance on cloud processing.

From an ethical perspective, managing user expectations is equally important. The system should not replace the user’s agency or the role of a human guide but rather serve as a supportive cognitive aid. Consequently, feedback should express probabilistic confidence—for instance, “probably stairs ahead”—rather than absolute statements, particularly when detections occur at lower confidence levels.

Contemporary discussions in AI ethics (UNESCO, 2023; IEEE, 2023) emphasise transparency and user autonomy. Visually impaired users should have access to clear information about how the system operates, including which object classes are being detected, the confidence thresholds applied, and how historical feedback can be reviewed or adjusted. Such openness not only promotes trust and accountability but also aligns assistive MR systems with emerging standards for ethical and human-centred artificial intelligence.

4.4 Study Limitations

In the present comparison, a complete characterisation of end-to-end latency (camera → detection → interface decision → TTS/haptic output) was not reported, nor was a formal evaluation of frames per second (FPS) conducted under varying lighting conditions and scene complexity. Furthermore, user studies involving blind and visually impaired (BVI) participants were not performed, such as A/B testing with different confidence thresholds or communication styles.

The dataset used in this study originated from a single building, which limits environmental diversity. Future work should expand the dataset to include various architectural layouts, different times of day, crowded environments, and unconventional obstacles.

Finally, the miss rate for safety-critical classes under fixed false-alarm budgets was not measured. Such a metric would more accurately reflect real-world risk in Mixed Reality navigation systems than aggregate indicators such as overall mAP.

4.5 Future Research Directions

Integrating object detection with depth estimation and Simultaneous Localisation and Mapping (SLAM) will enable a transition from simple binary recognition — “object present or absent” — to a spatially aware understanding of “where and how far.”

In practice, this advancement would allow the system to deliver context-rich feedback such as “stairs ahead, approximately three meters away, two steps down,” significantly enhancing the functional value of assistive navigation.

Future iterations should also incorporate adaptive confidence thresholds learned from user behaviour. For example, if a user frequently responds to low-confidence detections and confirms their relevance, the system could gradually lower its operational threshold. Conversely, if such cues are routinely ignored, the threshold could automatically increase.

This user-specific calibration would create a personalised balance between sensitivity and precision, improving both usability and trust.

Another essential step involves measuring and optimising end-to-end latency, decomposed into individual stages, image acquisition, inference, decision logic, and speech or haptic synthesis, to guarantee a hard upper bound on response delay during critical scenarios (e.g., a rapid approach to an edge or obstacle).

At the model level, an ensemble approach may prove beneficial: employing YOLOv9-t as the baseline high-recall detector, and selectively reinforcing it with YOLOv8 for classes where precision is paramount. Additional promising directions include continual learning and knowledge distillation techniques, which would enable fast model adaptation to new indoor environments using small sample sets, without the need for complete retraining.

Finally, user-centred evaluation should form the cornerstone of future studies, combining both subjective metrics—such as user trust, cognitive load, and interface learnability—and objective indicators, including reaction time and the number of navigation errors in controlled tasks (e.g., moving from point A to point B). Together, these assessments will provide a more holistic understanding of system performance and its real-world assistive potential.

5. Conclusions

The comparison between YOLOv8 and YOLOv9-t demonstrates that, under realistic conditions—comprising 17 object classes, small object sizes, and variable lighting—both architectures achieve similarly high overall performance in terms of mAP@0.5 and F1-score.

The decisive factor for deployment in Mixed Reality (MR) systems for visually impaired users, however, lies in the models' behaviour at lower confidence thresholds and their respective convergence speeds. YOLOv9-t reaches its optimal F1 value at approximately 0.40, which facilitates calibration toward higher recall, and it converges faster (after 60 epochs versus 100 for YOLOv8), supporting more frequent model updates and continuous learning.

The reduced number of semantic confusions between visually similar object classes, combined with shorter inference times, further enhances the practical utility of YOLOv9-t. For these reasons, the model is recommended as the default detector for MR/VR assistive systems supporting blind and visually impaired users, with YOLOv8 optionally employed as a precision-refinement layer for selected high-confidence categories.

Implementation within Unity Passthrough and ONNX Runtime Mobile on the Meta Quest 3 enables fully on-device inference, thereby strengthening user privacy and reducing system latency and operational costs.

In summary, the YOLOv8 and YOLOv9-t architectures represent a significant advancement toward the development of more intelligent, ethical, and accessible assistive systems for visually impaired individuals. Their integration into mixed-reality environments—when guided by principles of inclusive and human-centred design—opens new possibilities in rehabilitation, education, and independent daily living.

In the coming years, such technologies are expected to become an integral component of smart buildings and public spaces, where artificial intelligence not only responds to human presence but also becomes a true partner in navigation, safety, and spatial awareness.

References:

Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934. Available at: https://arxiv.org/abs/2004.10934.

Fogli, D., Torsi, S., Persico, S. 2025. Disability inclusion in extended reality (XR) research: A systematic review. Information, Communication & Society. Doi: 10.1080/1369118X.2025.2535426. Available at: https://www.tandfonline.com/doi/full/10.1080/1369118X.2025.2535426.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S. and Murphy, K. 2017. Speed-Accuracy Trade-offs for Modern Convolutional Object Detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3296–3297. doi: 10.1109/CVPR.2017.351. Available at: https://arxiv.org/abs/1611.10012.

IEEE. 2023. Ethically Aligned Design: A Vision for Prioritising Human Well-being with Autonomous and Intelligent Systems. Version 1. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v1.pdf.

Jocher, G., Chaurasia, A., Qiu, J. 2024. Ultralytics YOLOv8: Model Overview & Docs. Ultralytics Documentation. Available at: https://docs.ultralytics.com/models/yolov8/.

Kim, J., Choi, S., Lee, J., Kim, Y. 2022. DeepNAVI: A deep learning based smartphone navigation assistant for blind people. Expert Systems with Applications, 212, p. 118720. doi: 10.1016/j.eswa.2022.118720. Available at: https://www.sciencedirect.com/science/article/pii/S0957417422017432.

Lee, H., Chen, Y., Jin, S., Liu, Y., Tan, P. 2025. VRSight: An AI-Driven Scene Description System to Improve Virtual Reality Accessibility for Blind and Low-Vision Users. arXiv preprint arXiv:2508.02958. Available at: https://arxiv.org/abs/2508.02958.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. 2014. Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T. (eds) Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8693. Cham: Springer. Doi: 10.1007/978-3-319-10602-1_48. Available at: https://arxiv.org/abs/1405.0312.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. 2016. SSD:

Single Shot MultiBox Detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol 9905. Cham: Springer. Doi: 10.1007/978-3-319-46448-0_2. Available at: https://arxiv.org/abs/1512.02325.

Park, J., Kim, H., Lee, S. 2025. YOLOv8-Based XR Smart Glasses Mobility Assistive System for Aiding Outdoor Walking of Visually Impaired Individuals in South Korea. Electronics, 14(3), p. 425. doi: 10.3390/electronics14030425. Available at: https://www.mdpi.com/2079-9292/14/3/425.

Qualcomm. 2024. Snapdragon XR2 Gen 2 Platform: Technical Overview. Qualcomm Technologies, Inc. Available at: https://docs.qualcomm.com/bundle/publicresource/87-73689-1_REV_A_Snapdragon_XR2_Gen_2_Platform_Product_Brief.pdf.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. doi: 10.1109/CVPR.2016.91. Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf.

Ren, S., He, K., Girshick, R., Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), pp. 1137-1149. doi: 10.1109/TPAMI.2016.2577031. Available at: https://arxiv.org/abs/1506.01497.

Tapu, R., Mocanu, B., Zaharia, T. 2020. Wearable assistive devices for the visually impaired: A state of the art survey. Pattern Recognition Letters, 137, pp. 37-52. doi: 10.1016/j.patrec.2020.05.031. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0167865518308602.

UNESCO. 2023. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO. Available at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

Wang, A., Chen, H., Liu, C., Chen, K., Xu, Z., Chen, Y. and Wang, C. 2024. YOLOv10: Real-Time End-to-End Object Detection. arXiv preprint arXiv:2405.14458. Available at: https://arxiv.org/abs/2405.14458.

Wang, C.Y., Yeh, I.H., Liao, H.Y.M. 2024. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. arXiv preprint arXiv:2402.13616. Available at: https://arxiv.org/abs/2402.13616.