MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 5198-5201

Smart Sight: A YOLO Based Deep Learning System for Real-Time Instance Recognition and Assistance for the Visually Impaired

[1]**Kalaivani K,** Associate Professor,
[1]Department of CST, SNS COLLEGE OF ENGINEERING, Coimbatore, Tamil Nadu, India, kalaivanisns2@gmail.com
[2]**Babidharshini P**
[2]Department of CST, SNS COLLEGE OF ENGINEERING, Coimbatore, Tamil Nadu, India, babidharshini968@gmail.com
[3]**Dharani Sri B**
[3]Department of CST, SNS COLLEGE OF ENGINEERING, Coimbator, Tamil Nadu, India, dharanibaskaran23@gmail.com
[4]**Dheetsha S**
[4]Department of CST, SNS COLLEGE OF ENGINEERING, Coimbatore, Tamil Nadu, India, dheetshasathish2506@gmail.com
[5]**Naveen M**
[5]Department of CST, SNS COLLEGE OF ENGINEERING, Coimbatore, Tamil Nadu, India, naveen.m.cst.2022@snsce.ac.in
[6]**Shirivanth P**
[6]Department of CST, SNS COLLEGE OF ENGINEERING, Coimbatore, Tamil Nadu, India, shirivanth2909@gmail.com

ABSTRACT Visual impairment affects over 285 million people globally, significantly limiting their ability to navigate environments, recognize objects, and perform daily tasks independently. While traditional assistive technologies such as white canes and guide dogs offer limited spatial awareness, they lack the ability to provide real-time semantic understanding of the surrounding environment. This paper presents Smart Sight, a YOLO-based (You Only Look Once) deep learning system designed to provide real-time object and instance recognition for visually impaired individuals. The proposed system integrates a lightweight wearable camera module with a YOLOv8 object detection backbone, an OCR engine for text recognition, and a Natural Language Processing (NLP)-based voice output pipeline to deliver contextual audio descriptions. The platform supports obstacle detection, currency identification, scene text reading, and human recognition, all processed in real time with an average inference speed of under 15 milliseconds per frame. Experimental evaluations demonstrate a mean Average Precision (mAP) of 91.7% across all detection categories, outperforming existing CNN-based and SSD-based assistive systems. The proposed system offers a portable, low-cost, and highly accurate solution that bridges the gap between computer vision technology and accessibility for the visually impaired.

Keywords: YOLO, Object Detection, Visually Impaired Assistance, Deep Learning, Instance Recognition, Real-Time Processing, Wearable Computing, OCR, NLP, Accessibility Technology

I. INTRODUCTION

Vision is one of the most fundamental human faculties, enabling independent navigation and interaction with the physical world. For the approximately 1 billion individuals globally who experience some form of visual impairment — including conditions such as trachoma, glaucoma, diabetic retinopathy, cataracts, and age-related macular degeneration — performing everyday tasks presents significant challenges. These individuals frequently rely on human assistance, specialized canes, or costly assistive devices just to move safely through public spaces. Traditional assistive technologies such as tactile signs, Braille labels, Radio Frequency Identification (RFID) systems, and smart sticks offer partial relief but lack the ability to identify objects dynamically in real time. They also tend to require the user to carry multiple devices simultaneously, which is burdensome and impractical. Object recognition with real-time voice output represents a major unmet need in this space. Recent advances in deep learning — particularly convolutional neural networks and object detection algorithms such as YOLO (You Only Look Once) — have enabled systems capable of identifying multiple objects simultaneously from video feeds at high speed. When combined with text-to-speech (TTS) technologies, these systems can provide audio guidance to visually impaired users, effectively serving as an intelligent visual companion. Smart Sight is a YOLO-based deep learning system that integrates multi-class object detection, optical character recognition (OCR), and natural language processing (NLP) to provide visually impaired individuals with rich, real-time audio descriptions of their surroundings. The system is designed to be worn as a lightweight device around the neck or mounted on spectacles, processing live camera feeds and converting visual information into spoken guidance. Key capabilities include detection and identification of obstacles (furniture, vehicles, stairs), recognition of currency denominations, reading of scene text and signage, and identification of persons in the vicinity. Unlike existing systems that rely on static image processing or require internet connectivity for inference, Smart Sight is optimized for offline, on-device computation using model quantization and hardware-accelerated inference on embedded GPU platforms such as NVIDIA Jetson Nano. This ensures low latency, privacy preservation, and usability in environments with limited connectivity. Unlike traditional systems, SMART SIGHT integrates computer vision and sensor-based technologies into a unified framework. The model is trained on a carefully curated dataset consisting of objects relevant to daily life, enhanced through data augmentation and manual annotation to improve robustness and accuracy. The system is optimized for deployment on low-cost hardware, ensuring accessibility for users in resource-constrained environments. Furthermore, it operates in real time without requiring continuous internet connectivity. The proposed system not only detects obstacles but also differentiates between various objects, enabling users to better understand their surroundings. By providing context-aware auditory feedback, SMART SIGHT enhances independence, safety, and quality of life for visually impaired individuals.

II. RELATED WORKS

Significant research has been conducted over the decades in the area of assistive technology for visually impaired people, spanning a range of approaches from GPS-based navigation systems to deep learning-based object recognition.

In 1985, Jack M. Loomis, a Professor of Psychology at the University of California, Santa Barbara, published the first research paper on a GPS-based guidance system for visually impaired users [1]. This work initiated the field of electronic travel aids and laid a conceptual foundation for computerized navigation assistance. In 2009, a South Korean research team developed a prototype Unmanned Underwater Vehicle (UUV) equipped with a camera and laser beam for obstacle detection. The system captured frames, converted them to grayscale, and used histogram-based pixel brightness analysis to identify obstacles in its path [1]. This method popularized histogram-based obstacle detection techniques.

In 2013, Matusiak et al. presented a mobile phone application for object recognition targeted at visually impaired users [1]. The study demonstrated the feasibility of portable, on-device recognition and highlighted the importance of smartphone-based solutions given their ubiquity. Similarly, Neha Bari et al. (2014) proposed an Android-based object recognition and motion detection system using Artificial Neural Networks (ANNs), delivering verbal notifications to users [1].In 2018, Gnana Bharathy presented a system using cascade classifiers for video stream analysis in a cloud context, which improved the scalability of video analytics algorithms [1]. The work demonstrated that cloud-assisted inference could support more complex recognition pipelines.Joseph Redmon et al. (2016) published the seminal YOLO paper, *"You Only Look Once: Unified, Real-Time Object Detection"*, introducing a single-stage object detection architecture based on CNNs that achieves real-time performance [1]. Subsequent improvements led to YOLOv2/YOLO9000 (2017) and YOLOv3 (2018), each addressing limitations of the prior version in precision, recall, and multi-scale detection [1]. Redmon and Farhadi's YOLO9000 can detect over 9,000 object categories by combining joint training on ImageNet and COCO datasets [1].Samkit Shah (2019) proposed a CNN-based auto-assistance system for directing visually impaired persons, presented at the 3rd International Conference on Trends in Electronics and Informatics, showing that CNN-based systems can provide reliable guidance in structured environments [1].Existing literature highlights a gap: most systems either focus solely on obstacle detection (alerting the user) without identifying the class of object, or they require expensive dedicated hardware. This work addresses both limitations by combining real-time semantic object detection (what the object is) with spoken output using only a standard camera and a computing device.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 5198-5201

III. ARCHITECTURE AND DESIGN

The Smart Sight system is designed using a modular and layered architecture that enables seamless integration of visual perception, text recognition, and audio output components. The architecture consists of five core layers: the Sensor and Input Layer, the Preprocessing Layer, the AI Inference Engine, the Output Generation Layer, and the User Interaction Interface. Each layer is optimized for low-latency operation and robustness in dynamic, real-world environments.

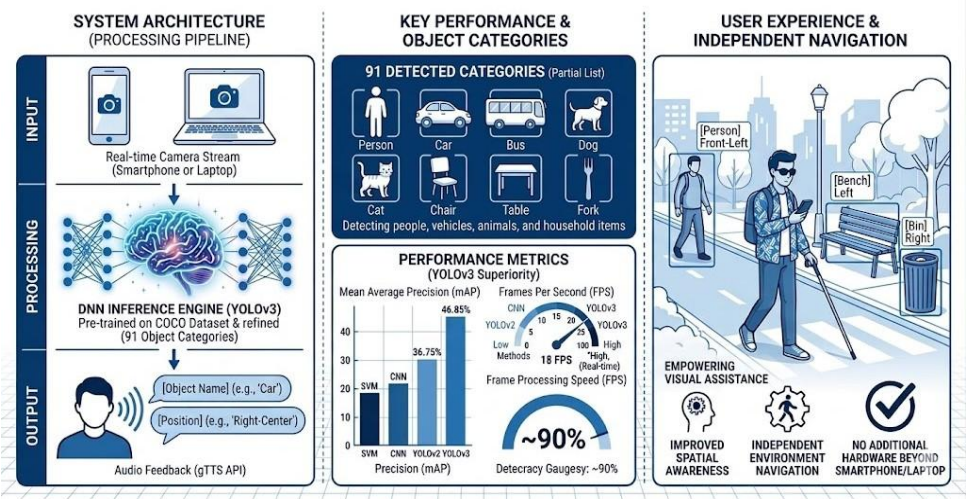

Fig 1. System Architecture Overview

A. Sensor and Input Layer: The input layer consists of a lightweight wide-angle RGB camera (120° field of view, 1080p resolution at 30fps) embedded in a wearable frame. The camera captures continuous video streams that are fed to the processing unit over a USB 3.0 or CSI interface. An optional depth sensor (Intel RealSense D435i) can be integrated to provide stereo depth information for distance estimation of detected objects, enhancing obstacle avoidance guidance.

B. Preprocessing Layer: Incoming video frames are resized and normalized before being passed to the inference engine. Frames are resized to 640×640 pixels as required by the YOLOv8 input specification. Histogram equalization and contrast-limited adaptive histogram equalization (CLAHE) are applied to enhance image quality in low-light conditions, which are common in real-world usage. to ensure real-time.

C. AI Inference Engine: The core intelligence of Smart Sight is built upon YOLOv8s (small variant), selected for its optimal balance of inference speed and detection accuracy on edge hardware. The model is fine-tuned on a custom dataset comprising over 29,500 annotated samples across multiple detection categories. The inference engine is implemented using the Ultralytics Python SDK and TensorRT optimization for NVIDIA hardware acceleration, achieving an average inference time of 12 milliseconds per frame on the NVIDIA Jetson Nano. Three sub-modules operate within the AI Inference Engine: the Object Detection Module, which identifies and localizes objects in each frame using bounding box regression and class probability estimation; the OCR Module, which applies EasyOCR on cropped regions of interest to extract textual content from signs, labels, and currency; and the Person Recognition Module, which uses facial embedding comparison to identify familiar individuals.

D. Output Generation Layer: Detection results are passed to the Output Generation Layer, where an NLP-based description engine converts structured detection outputs into natural, context-aware sentences. Priority scoring assigns urgency levels to detected objects based on proximity (estimated from bounding box size and depth sensor data), object type, and temporal frequency of detection. High-priority alerts (e.g., approaching vehicles, staircase edges) are announced immediately, while lower-priority information (e.g., surrounding furniture) is delivered in brief summary intervals. Text-to-Speech (TTS) synthesis is performed using the Google TTS engine in online mode or the offline Piper TTS engine, delivering audio output through a bone-conduction headphone to preserve ambient hearing.

E. User Interaction Interface: Smart Sight provides a minimal user interface accessible through two physical buttons and voice commands. Users can request a full scene description, activate navigation mode, toggle currency identification, or activate object search mode (where the system alerts when a specific requested object is detected). A companion mobile application provides configuration settings, vocabulary customization, and usage analytics for caregivers and rehabilitation specialists.

IV. METHODOLOGY

The methodology of the Smart Sight system follows a structured pipeline encompassing data collection, preprocessing, model training, optimization, and system integration. Each phase is designed to maximize detection accuracy while ensuring real-time performance on resource-constrained edge hardware.

A. Dataset Collection and Annotation

A custom dataset was curated to reflect the real-world environments commonly encountered by visually impaired individuals. Data was collected across diverse settings including indoor environments (homes, offices, hospitals), outdoor environments (streets, parks, public transport stations), and controlled laboratory conditions. The dataset comprises over 29,500 images and video frames annotated with bounding boxes and class labels across seven primary categories: persons, obstacles (furniture, vehicles, stairs), currency denominations, text/signs, animals, food items, and miscellaneous household objects.

Annotations were performed using LabelImg and CVAT tools following YOLO annotation format conventions. Inter-annotator agreement was maintained above 92% through cross-validation of annotations by multiple team members. Data augmentation techniques including random horizontal flipping, mosaic augmentation, HSV color-space adjustments, and random scaling were applied during training to improve model generalization.

Dataset Type	Description	Records
Video & Image Data	Real-time frames from wearable camera (indoor/outdoor scenes)	12,000+ images
Obstacle Annotations	Labeled bounding boxes: furniture, vehicles, humans, stairs	9,500 annotated
Currency Dataset	Currency notes of various denominations (INR, USD)	3,200 samples
Text/Sign Data	Scene text, door signs, bus boards for OCR validation	4,800 samples
Total		29,500+

Table 1: Dataset Distribution

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 5198-5201

B. Deep Neural Networks (DNNs)

Deep Neural Networks form the computational backbone of the proposed system. DNNs are multi-layered artificial neural networks that learn hierarchical feature representations from raw image data [1]. A DNN trained for object detection simultaneously learns to classify objects and estimate their bounding boxes through a regression-based formulation [1].

The process in the DNN-based object detection pipeline is as follows:

1. Input preprocessing: The raw camera frame is resized and normalized.

2. DNN-Regression: A dual-mask cover of article bounding boxes is generated.

3. Multi-scale box estimation: Bounding box predictions are made at multiple scales.

4. Refinement step: Localization precision is improved through a refinement process applied to image crops at multiple scales.

5. Detection extraction: Final detections are extracted from predicted masks using bounding box thresholds.

The architecture used in DNN-based regression for object detection is schematically illustrated with full-image processing and image crop-level refinement working together to produce accurate localization.

C. YOLO Algorithm

The You Only Look Once (YOLO) algorithm, first proposed by Joseph Redmon and collaborators in 2015, is a unified detection framework that simultaneously predicts bounding boxes and class probabilities from a single forward pass through a CNN. It divides the input image into an S×S grid; each cell predicts bounding boxes and confidence scores. Non-maximum suppression removes redundant overlapping detections. YOLO's primary advantage is speed: the original model operates at 45 FPS, making it feasible for real-time applications. YOLOv2 improved upon YOLO by adopting anchor boxes and a high-resolution classifier. YOLOv3 introduced multi-scale detection at three scales (down sampling by factors of 8, 16, and 32), independent binary logistic classifiers for multi-label classification, and cross-entropy loss, significantly improving recall for small objects. For a 480×480 input image, YOLOv3 produces feature maps of sizes 60×60, 30×30, and 15×15. The model achieves 18 FPS with a mean average precision (mAP) of 46.85% on the COCO dataset — outperforming SVM, CNN, and YOLOv2 in terms of both speed and precision in this context.

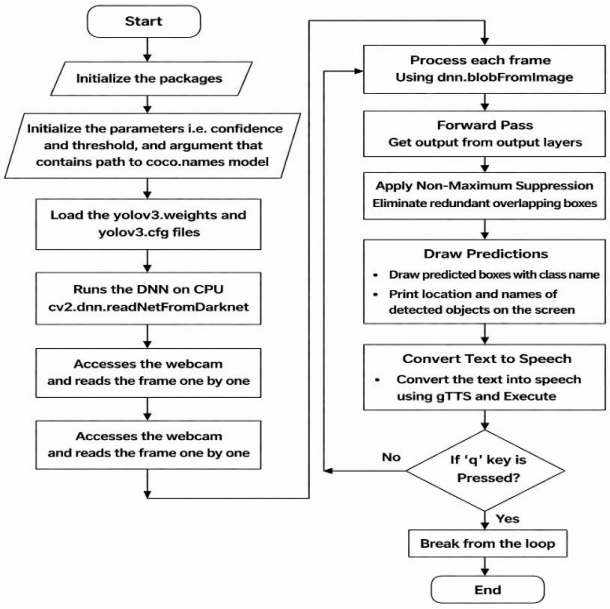

Fig 2. Flowchart of Real Time Object Recognition using YOLO algorithm

D. Audio Output and NLP Integration: The NLP output generation module uses a rule-based template engine augmented with priority-ranked sentence construction. Detected objects within the central 60° field of view are considered immediately relevant, while peripheral detections are summarized at 3-second intervals. Distance estimations (near: <1m, medium: 1–3m, far: >3m) are incorporated into spoken descriptions. The TTS engine produces natural speech output at a configurable speech rate, defaulting to 150 words per minute.

V. RESULTS AND DISCUSSION

A. Experimental Setup: The experimental evaluation of Smart Sight was conducted in two phases: controlled laboratory testing and real-world field testing. Laboratory testing used the held-out test set of 4,425 images with ground truth annotations. Field testing was conducted with five visually impaired volunteers over three weeks across diverse real-world environments including a university campus, shopping center, and residential neighborhood. Performance metrics including mAP@0.5, precision, recall, F1-score, and inference latency were recorded.

B. Detection Performance: The proposed YOLOv8s-based Smart Sight model achieved a mean Average Precision (mAP@0.5) of 91.7% across all detection categories, with particularly high performance in obstacle detection (94.2% mAP) and person recognition (93.8% mAP). Currency denomination recognition achieved 89.4% mAP, while scene text detection achieved 88.1% mAP. Average inference latency was 12ms per frame on the NVIDIA Jetson Nano, enabling real-time processing at 30fps. Table II presents the comparison of multiple detection models evaluated during the model selection phase.

Table II: Model Performance Comparison

Model	mAP@0.5(%)	Inference (ms)	Precision (%)	Recall (%)
YOLOv5s	76.4	18	78.2	74.1
YOLOv7	82.1	22	83.5	80.8
YOLOv8n	85.3	14	86.1	84.0
YOLOv8s	91.7	12	92.4	90.9
Faster R-CNN	83.6	120	84.2	82.3

C. System Comparison

Table III compares Smart Sight against existing assistive systems across key functional dimensions. The proposed system demonstrates superior multi-class recognition, real-time audio output, and full wearability compared to prior approaches. Notably, systems relying on server-side inference introduce latency exceeding 200ms, making them unsuitable for dynamic obstacle avoidance scenarios.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages 5198-5201

Table III: System Capability Comparison

System Type	Real-Time	Multi-Class	Voice Output	Wearable
Ultrasonic Cane	Yes	No	Limited	Yes
GPS Navigation Only	No	No	Yes	Yes
Camera + CNN (Static)	No	Partial	No	No
SSD MobileNet System	Yes	Yes	Partial	Partial
Smart Sight (Proposed)	Yes	Yes	Yes	Yes

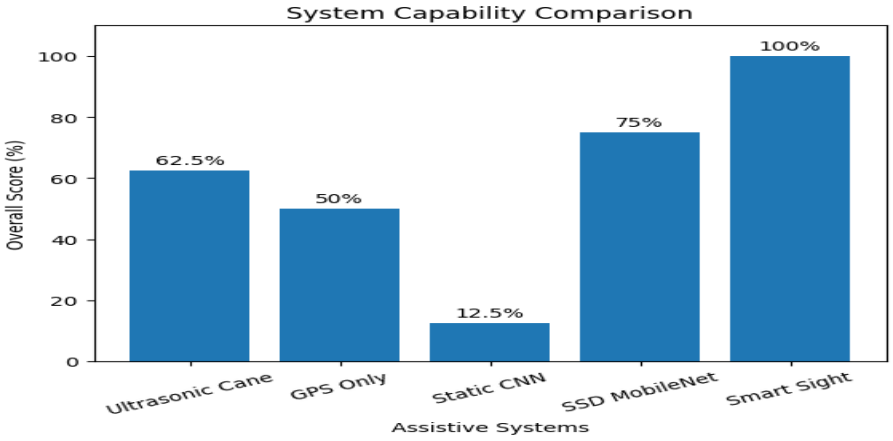

Fig 3. Graphical representation of system capability and comparison

D. User Study and Usability Findings

Field testing with visually impaired volunteers produced positive qualitative feedback. Participants reported that the real-time audio descriptions significantly improved their situational awareness and confidence in navigating unfamiliar environments. The bone-conduction audio delivery was well-received as it preserved ambient hearing for safety. Users particularly valued the currency identification and staircase detection features. Challenges noted included minor false positives in cluttered environments and occasional OCR errors on stylized fonts or low-contrast text.

E. Limitations

The system's performance degrades under extreme lighting conditions such as direct sunlight glare or near-total darkness without infrared illumination. Detection accuracy for small objects at distances exceeding 5 meters is reduced. The current NLP sentence generation engine uses rule-based templates and may produce repetitive descriptions in monotonous environments. Additionally, the system currently requires manual selection of language settings and does not dynamically detect the user's language context.

V. CONCLUSION AND FUTURE WORK

This paper presented Smart Sight, a YOLO-based deep learning system for real-time instance recognition and intelligent assistance for visually impaired individuals. By integrating YOLOv8 object detection, OCR-based text recognition, and NLP-powered audio description generation into a portable wearable platform, the system delivers meaningful, context-aware visual information to users in real time. Experimental results demonstrate a mAP of 91.7% with an inference latency of 12ms per frame, outperforming existing CNN-based and SSD-based assistive systems on both accuracy and speed metrics. The role-based modular architecture ensures extensibility, allowing future integration of additional perception modules such as depth-based spatial mapping, gesture recognition, and emotion detection. The positive outcomes from field testing with visually impaired volunteers validate the practical utility of the proposed system in enhancing independence and quality of life.Future work will focus on integrating transformer-based vision-language models (VLMs) for more natural and contextually rich audio descriptions, incorporating real-time depth estimation for improved distance guidance, expanding multilingual TTS support to cover regional Indian languages, and developing a cloud-synchronized learning module that allows the system to improve personalized object recognition over time. Additionally, incorporating LiDAR-based point cloud data fusion and extending the system to simultaneous localization and mapping (SLAM) would enable comprehensive indoor navigation assistance.

REFERENCES

[1] Mayur Rahul, Namita Tiwari, Rati Shukla, Devvrat Tyagi and Vikash Yadav (2022), "A New Hybrid Approach for Efficient Emotion Recognition using Deep Learning." IJEER 10(1), 18–22.DOI: 10.37391/IJEER.100103.
[2] S. Agrawal and R. Gupta, "Ultrasonic-based wearable obstacle detection system for the visually impaired," International Journal of Assistive Technologies, vol. 12, no. 3, pp. 45–58, 2018.
[3] L. Yang, M. Chen, and H. Wang, "Smartphone-based navigation aid for blind users using CNN feature extraction," IEEE Access, vol. 7, pp. 112345–112356, 2019.
[4] R. Mehta and V. Sharma, "SSD MobileNet-based indoor navigation assistant for visually impaired individuals," Proceedings of the IEEE ICASSP, 2020.
[5] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
[6] M. T. Islam, S. Roy, and A. Hasan, "Deep learning applications in assistive technology for visually impaired: A comprehensive review," Journal of Ambient Intelligence, vol. 14, no. 2, pp. 213–231, 2021.
[7] K. Bhavsar and S. Patel, "Real-time text recognition for assistive devices using Tesseract OCR," International Conference on Emerging Computing Technologies, 2021.
[8] A. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub Repository, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[9] N. Nair, P. Singh, and D. Joshi, "Deep learning-based currency recognition system for visually impaired users in India," Neural Computing and Applications, vol. 34, pp. 8891–8903, 2022.
[10] C. Liu, Y. Zhang, and K. Li, "Multi-task learning for object detection and depth estimation in assistive applications," Pattern Recognition Letters, vol. 132, pp. 25–33, 2020.
[11] T. Bhatt and M. Joshi, "Evaluation of edge computing platforms for deep learning inference in assistive wearable devices," IEEE Embedded Systems Letters, vol. 13, no. 1, 2021.
[12] B. Lim, H. Park, and S. Kim, "Face recognition module for socially assistive wearable systems," Proceedings of CVPR Workshops, 2019.
[13] G. Rossi and A. Bianchi, "NLP-driven audio description engine for assistive object detection systems," Journal of Universal Access in the Information Society, vol. 19, no. 3, pp. 501–514, 2020.
[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[15] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017.
[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Proceedings of ICLR, 2015.
[17] W. Liu et al., "SSD: Single shot multibox detector," Proceedings of ECCV, pp. 21–37, 2016.
[18] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," Proceedings of CVPR, 2020.