Cluster Computing (2025) 28:997
https://doi.org/10.1007/s10586-025-05740-z

Enhanced obstacle detection using bilateral vision-aided transformer neural network for visually impaired persons

Ala Alarood[1,2] · Mohammed Salem Atoum[3] · Azizah Abdul Manaf[4,5] · Adamu Abubakar[2,4,5] · Izzat Alsmadi[6,7]

Received: 8 June 2025 / Revised: 1 August 2025 / Accepted: 25 August 2025

Abstract
Obstacle detection remains vital in autonomous navigation and assistive technologies, especially for visually impaired individuals. This work introduces an enhanced obstacle detection framework based on a Bilateral Vision Transformer and Convolution Kernel Neural Network (BViT-CKNN). The system incorporates stereo vision data and applies a bilateral filter to reduce noise while preserving edge details. A Vision Transformer (ViT) model is then used for global feature extraction, and a Convolution Kernel Neural Network (CKNN) captures fine-grained local features. Evaluated using the COCO dataset, the proposed BViT-CKNN achieves superior performance in precision (0.93), recall (0.91), F1-score (0.92), and Mean Absolute Error (MAE) reduction (3.16%) compared to existing methods.

Keywords Visually impaired · Obstacle detection · Bilateral filter vision transformer · Convolution Kernel Neural Network

1 Introduction

Precise obstacle detection is crucial in various technology-driven fields, such as autonomous vehicles, mobile robotics, and surveillance systems. The significance of robust and dependable obstacle detection in assistive technology for the visually impaired is paramount. Individuals with visual impairments encounter substantial difficulties traversing known and unknown surroundings due to the lack of visual indicators. Thus, technological devices that can detect and identify barriers in real-time significantly improve autonomy, safety, and quality of life. Nonetheless, despite progress in computer vision and artificial intelligence, creating a universally reliable system for real-time obstacle identification continues to be a formidable challenge owing to the diversity and intricacy of real-world settings.

Conventional obstacle identification systems frequently depend on manually designed features or sensor inputs (e.g., LiDAR, ultrasonic sensors, or stereo cameras), which may constrain range, ambient illumination conditions, and surface characteristics. Although these strategies have certain advantages, they are inadequate in dynamically changing or unstructured surroundings, prevalent in outdoor and urban areas where visually impaired individuals often traverse. The growing accessibility of computer resources and extensively annotated datasets has elevated deep learning

✉ Ala Alarood
aasoleman@uj.edu.sa

Mohammed Salem Atoum
m.atoum@ju.edu.jo

Azizah Abdul Manaf
azizahm@utar.edu.my

Adamu Abubakar
adamu@iium.edu.my

Izzat Alsmadi
ialsmadi@tamusa.edu

1 College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

2 King Salman Center For Disability Research, Riyadh, Saudi Arabia

3 Department of Computer Science, The University of Jordan, Amman, Jordan

4 Department of Computing, LKC Faculty of Engineering and Science, University. Tunku Abdul Rahman (UTAR) Sungai Long Campus, Perak, Malaysia

5 Department of Computer Science, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia

6 Department of Computing, Engineering and Mathematical Sciences, Texas A&M San Antonio, San Antonio, TX, USA

7 School of Information Technology and Systems, University of Jordan, Aqaba, Jordan

Published online: 15 October 2025

methods, providing more scalable and flexible solutions. Convolutional Neural Networks (CNNs) have demonstrated significant efficacy owing to their capacity to learn spatial hierarchies and local properties from unprocessed picture data. Concurrently, transformer-based architectures—initially designed for natural language processing—have started to exhibit remarkable efficacy in computer vision problems, primarily owing to their ability to simulate long-range relationships and global context via self-attention processes.

In recent years, object detection algorithms such as YOLO (You Only Look Once) [1] and MobileNet [2] have garnered considerable interest for their capacity to execute rapid and efficient object identification in real-time applications. YOLO models partition an input image into grids and directly forecast bounding boxes and class probabilities from whole images in a single evaluation, yielding rapid inference. Conversely, Mobile Net prioritizes lightweight topologies for mobile and embedded devices, including depth wise separable convolutions, to minimize computational expense. Notwithstanding their benefits, both models face performance compromises when implemented in intricate real-world settings characterized by several object categories, fluctuating illumination, occlusions, and diminutive objects. YOLO may encounter difficulties with precise localization in densely cluttered environments, whereas Mobile Net may sacrifice precision for efficiency. These restrictions are particularly significant when the end-user is a visually impaired individual who needs a high-reliability level for safe and efficient navigation. An in-depth qualitative and quantitative analysis employing artificial intelligence techniques was investigated in [3]. A low-cost, single device mechanism provided with obstacle detection and identification features to improve user navigation without employing multiple detection devices was presented in [4]. Yet another convolutional transformer to improve efficiency of multi-head attention with improved with improved accuracy was proposed in [5] to assist visually challenged. In [6] a hybridization of YOLOv4 and the COCO dataset, facilitating to improve object recognition while controlling the advantages of obstacle recognition was proposed with improved accuracy.

Blindness relates to the visual perception loss that can bring about mobility and self-reliance problems for visually impaired people. As a consequence some research has been performed by several researchers with the intent of solving problems encountered in day to day life. Nevertheless, people with visual impairments still pose numerous issues that make their lives painful. Machine learning technique was applied in [7] for obstacle avoided with minimal time.

Computer vision techniques were applied in [8] with the intent of detecting objects in the nearly areas and convey this to the visually impaired uses via voice messages accurately. Yet another method to concentrate on accuracy aspects of obstacle detection, employing lightweight feature aware enhanced target detection network was proposed in [9]. Artificial Intelligence (AI) techniques were integrated with sensoring mechanism [10] for visually impaired people. An optimized deep learning algorithm employing social optimization technique based on the Deep Convolutional Network (Deep CNN) for ease facilitation of visually damaged persons was proposed in [11]. By employing this Deep CNN attained improved accuracy. Yet another enhanced feature attention method to focus on the reproducibility was designed in [12] with improved precision.

To address this gap, we propose new hybrid architecture named Bilateral Vision Transformer with Convolutional K-Nearest Neighbor (BViT-CKNN). This model utilizes the synergistic advantages of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) to establish a more comprehensive and efficient obstacle detection pipeline. Vision Transformers excel at capturing long-range dependencies and global context via their self-attention mechanisms, which is particularly advantageous in situations when impediments manifest in many positions and shapes. Nonetheless, ViTs may encounter difficulties recognizing tiny local details because of the absence of inductive biases characteristic of CNNs. To mitigate this constraint, we incorporate a CKNN (Convolutional K-Nearest Neighbor) module to enhance the object boundary information and local features that may be compromised in the attention-based global analysis. By integrating both modules, BViT-CKNN can concurrently analyze global semantics and localized characteristics, yielding enhanced precision and reliability in obstacle recognition.

A crucial element of our system is the bidirectional filtering procedure utilized in the preprocessing phase. The bilateral filter is a non-linear, edge-preserving, noise-reducing smoothing filter that retains crisp edges by evaluating both the spatial proximity and intensity similarity of pixels. For those with visual impairments, preserving edge information is especially vital when minor misdetections may result in accidents. Implementing bilateral filtering before feature extraction enhances picture feature clarity, diminishes artifacts, and optimizes the efficacy of following detection phases.

The hybrid model operates via the subsequent pipeline: (1) input images undergo preprocessing through bilateral filtering to enhance edge definition and diminish visual noise; (2) a Vision Transformer module extracts elevated, global semantic features from the image; (3) these features are transmitted to a CKNN module that accentuates local detail refinement, facilitating precise delineation of object contours; (4) a classification head forecasts object

categories utilizing the amalgamated feature representations. This sequential and synergistic approach enhances detection accuracy and assures resilience against real-world visual aberrations, rendering it especially suitable for assistive applications.

We assess our suggested framework utilizing the COCO dataset, a prevalent benchmark in object detection encompassing a broad array of commonplace objects in various settings and orientations. Our findings indicate that BViT-CKNN exceeds contemporary methodologies in essential performance parameters, such as mean Average Precision (mAP), precision-recall equilibrium, and inference reliability across diverse image circumstances. These enhancements indicate that our methodology is efficient in controlled benchmark settings and shows potential for implementation in real-world assistive systems.

The key contributions of this study are outlined as follows:

- We introduce a novel hybrid deep learning architecture, BViT-CKNN, which combines the global contextual understanding of Vision Transformers with the localized feature refinement of CNN-based K-Nearest Neighbor (CKNN) operations to achieve improved obstacle detection.
- A bilateral filtering technique is applied during preprocessing to enhance edge sharpness and overall image quality, facilitating more accurate obstacle identification under varied visual conditions.
- The proposed model adopts a modular design, where the Vision Transformer captures long-range semantic dependencies while the CKNN module ensures fine-grained spatial precision and detail enhancement.
- Extensive experiments conducted on the COCO dataset show that our approach outperforms existing state-of-the-art object detection methods, delivering higher accuracy and greater robustness across diverse scenarios.

The remainder of this paper is organized as follows: Sect. 2 reviews existing literature and methodologies on obstacle detection using traditional filtering and deep learning techniques. Section 3 briefly introduces the proposed BViT-CKNN framework, including its architecture, components, and processing pipeline. Section 4 presents our experimental setup, dataset details, training procedures, and evaluation metrics. Section 5 comprehensively assesses the proposed method compared with other state-of-the-art techniques, analyzing performance under various conditions. Finally, Sect. 6 concludes the paper by discussing future directions and potential applications in real-world assistive systems.

2 Related work

Obstacle detection for visually impaired individuals has garnered significant attention due to advances in computer vision, AI, and deep learning, aiming to develop real-time, accurate, and context-aware assistive systems. Existing approaches primarily use modified object detectors like YOLOv5 and SSDLite MobileNetV2, offering fast inference and precision but often lacking robust error handling and generalizability. Transformer-based models and attention mechanisms have improved global context understanding and classification accuracy, yet struggle with local detail refinement and reliable error minimization. Transfer learning with pre-trained CNNs enhances accuracy and reduces training time but faces limitations in adaptability under varied conditions. Sensor-integrated solutions demonstrate promise in practical deployment but frequently compromise speed, accuracy, or user convenience. These limitations across methodologies underscore the need for a hybrid model like BViT-CKNN, which integrates global and local feature learning for enhanced obstacle detection.

A holistic review of the current state of the art methods in edge deep learning concentrating on computer vision applications, in specific medical diagnostics was investigated in [13]. An overview of the indispensable concepts and technical advantages of edge deep learning was presented, drawing attention to the potentiality of this technique in revolutionizing wide range of domains. A review of Artificial Neural Networks in ascertaining the requirements of visually impaired persons was designed in [14].

Numerous artificial intelligence techniques focusing on visually impaired was proposed in [15]. A Viola Jones and TensorFlow Object Detection method to design modest and versatile framework for visually impaired to assist them in their daily routines was designed in [16]. By using this method obstructions were detected with a notable high efficiency. Specific obstacle detection facing visually impaired in recognizing food was proposed in [17] employing multi-scale feature fusion network with texture feature extraction model. With this type of design ensured to classify food as accurately as possible.

Investigating discerning visual patterns from image local notable regions is extensively utilized for fine-grained visual classification tasks, to name a few being, classification plant or animal species. An extensive amount of complicated networks have been evolved for discerning learning feature representations. In [18] a novel local structure information (LSI) learning method was proposed to fine notably regions accurately.

An adaptable grid generated based on the immense object size within user's proximity employing innovative neural perception was designed in [19] with improved

accuracy rate. Also to extract rich feature on the immense object size within user's proximity detailed feature extraction model employing encoder decoder convolution model was proposed in [19] that in turn achieved improved accuracy. An assistive system employing point cloud registration was designed in [20] for obstacle detection. Novel strategy YOLOSEG was proposed in [21] for intelligent road segmentation and obstacle detection of railway trespasser. But precision was not improved. Lightweight dual-branch semantic segmentation network was developed in [22] to improve water surface obstacle detection. However failed to enhance the recall.

Several systematic reviews and meta-analyses have mapped the landscape of assistive technologies for the visually impaired. For example, in [23], a review of 54 studies highlighted significant advances in localization and mapping techniques, while [24] focused on computer vision methods. Despite the breadth of approaches, these studies often reveal persistent gaps, including inadequate performance in uncontrolled environments, lack of portability, and absence of real-time feedback mechanisms.

More recent approaches have proposed transformer-based models, such as in [25], targeting multi-object detection with improved classification accuracy. Others, like [26], used Obstacle-Transformer architectures to maintain constant inference time for trajectory forecasting. Despite promising results, many of these models fail to reduce MSE sufficiently and lack balanced local and global feature extraction—both critical for reliable obstacle detection.

A comparative summary of the key findings of previous research is presented in Table 1. The table highlights a range of deep learning approaches in recent studies to aid visually impaired individuals through obstacle detection systems. Modified YOLOv5, as used by Ahmed Ben Atitallah et al. [1], demonstrated improved speed and precision due to backbone enhancements but lacked Mean Absolute Error (MAE) considerations. Raihan Bin Islam et al. [2] applied SSDLite MobileNetV2, offering lightweight deployment and high precision, though it struggled in generalizing across diverse environments. Xinrong Li et al. [27]'s multimodal attention network showed strong classification accuracy by integrating spatial and temporal cues but did not directly address error minimization. Pre-trained CNN models and deep transfer learning approaches are enhanced usability and accuracy. However, they faced limitations in reliability and MAE reduction, as seen in the works of Wasiq Khan et al. [28] and Bineeth Kuriakose et al. [29], respectively. Cost-effective systems, such as the one by Xinnan Leong et al. [30], improved F1 scores in single-device setups but suffered from slow inference. Advanced architectures like convolutional transformers and hybrid YOLOv4-COCO models, explored by Sunnia Ikram et al. [25] and Yahia Said et al. [31], achieved strong accuracy and detection rates, although their error mitigation remained inadequate. Lastly, transformer-based detection, notably in the study by Nasrin Bayat et al. [32], provided robust class recognition but lacked contextual depth, underscoring the need for more holistic and error-aware solutions in assistive navigation technologies:

Table 1 Comparative summary of the previous research studies key findings

Study	Method	Contribution	Merits	Limitations
[1]	Modified YOLOv5	Enhanced backbone and training optimization for low-vision assistance	Improved speed and precision	MAE not addressed
[2]	SSDLite MobileNetV2	Lightweight detector for assistive tools	High precision	Poor generalization in varied scenarios
[27]	Multimodal attention network	Integrated spatial and temporal attention	High classification accuracy	No MAE minimization
[28]	Pre-trained CNN	Improved usability and scene perception	Enhanced accuracy	Low reliability
[29]	Deep transfer learning	Leveraged pre-trained models for accuracy	Good performance	Failed to reduce MAE
[30]	Low-cost single-device system	Focused on F1 score in affordable setups	Increased F1 score	Slow inference
[25]	Convolutional Transformer	Combined CNN and attention mechanisms	Strong accuracy	Inadequate error minimization
[31]	YOLOv4+COCO	Hybrid method for object detection	High detection rate	Still unreliable for visually impaired
[32]	Transformer-based obstacle detection	Focused on classification accuracy	Strong class recognition	Incomplete context understanding

3 Research gaps and motivation for BViT-CKNN

The collective review reveals a recurring pattern: while various models succeed in improving one or two performance aspects (accuracy, precision, speed), they often fail to deliver a comprehensive solution that balances local detail preservation, global context awareness, and error minimization—all of which are essential for the safe navigation of visually impaired individuals.

Additionally, several existing approaches neglect the integration of edge-preserving filters, such as bilateral filters, which can play a critical role in enhancing image clarity and object boundary detection, especially under variable lighting and environmental noise.

This motivates our proposed framework, Bilateral Vision Transformer with Convolutional K-Nearest Neighbor (BViT-CKNN). Unlike existing models, BViT-CKNN:

- Enhances input image quality using bilateral filtering to preserve edges and reduce noise.
- Combines a Vision Transformer (ViT) for global semantic understanding with a Convolutional KNN (CKNN) module for fine-grained local detail refinement.
- Maintains a low error margin while ensuring real-time performance.
- Demonstrates superior accuracy and robustness on challenging datasets such as COCO.

Our approach sets a new benchmark for reliable, efficient, and practical obstacle detection tailored for visually impaired individuals by addressing the local–global feature extraction gap and explicitly targeting error minimization metrics.

4 Methodology

Vision plays a critical role in human perception and daily activity navigation. For individuals with visual impairments, the absence or limitation of sight imposes substantial challenges in obstacle avoidance and environmental interaction. To address these challenges, this study proposes an advanced framework—Bilateral Vision Transformer and Convolution Kernel Neural Network (BViT-CKNN)—for enhanced obstacle detection.

The proposed method integrates a Bilateral Filter Vision Transformer for global contextual feature extraction and a Convolution Kernel Neural Network for localized, fine-grained feature refinement. The architecture is designed to capture high-level semantic information and detailed local characteristics, improving object classification and obstacle identification accuracy, especially in real-world, complex environments.

4.1 Architecture overview

The overall architecture of BViT-CKNN is illustrated in Fig. 1. The figure demonstrates the overall workflow of the proposed BViT-CKNN hybrid obstacle detection framework, which is designed to assist visually impaired individuals with enhanced accuracy and precision. The process begins with the COCO dataset, which contains 80 object classes commonly used for object detection tasks. These classes undergo bilateral filter-based preprocessing, which enhances image quality by preserving edges while reducing noise—crucial for effective obstacle recognition in diverse and real-world visual conditions.

Following preprocessing, the data is passed through two parallel yet complementary modules. The first is the Bilateral Filter Vision Transformer-based Global Feature Extraction, which captures long-range semantic relationships and contextual information across the entire image. The second is the Convolution Kernel Neural Network-based Local Feature Extraction, which focuses on fine-grained spatial details and precise boundary detection using localized convolutional operations.

The outputs of both modules are integrated to produce an accurate and precise obstacle detection result, benefiting from global context understanding and local feature refinement. This architecture effectively balances the strengths of transformer models in capturing broad contextual patterns and CNNs in enhancing spatial accuracy, making it highly suitable for assistive technologies in complex environments.

4.2 Experimental dataset

The experimental evaluation of the proposed BViT-CKNN model is conducted using the COCO (Common Objects in Context) dataset, a comprehensive and challenging benchmark extensively adopted for object detection, segmentation, and image captioning tasks. The COCO dataset is particularly well-suited for obstacle detection systems targeted at visually impaired individuals due to its diversity in scene contexts and object categories. It contains over 165,000 annotated images, with over 80 object categories, offering instance-level annotations and contextual information such as object segmentation masks and key points. These rich annotations enable the development and testing of models that detect objects and understand spatial relationships and complex environments—capabilities essential for real-time assistive navigation systems.

The dataset is systematically organized into multiple semantic categories, as shown in Table 2, which maps object

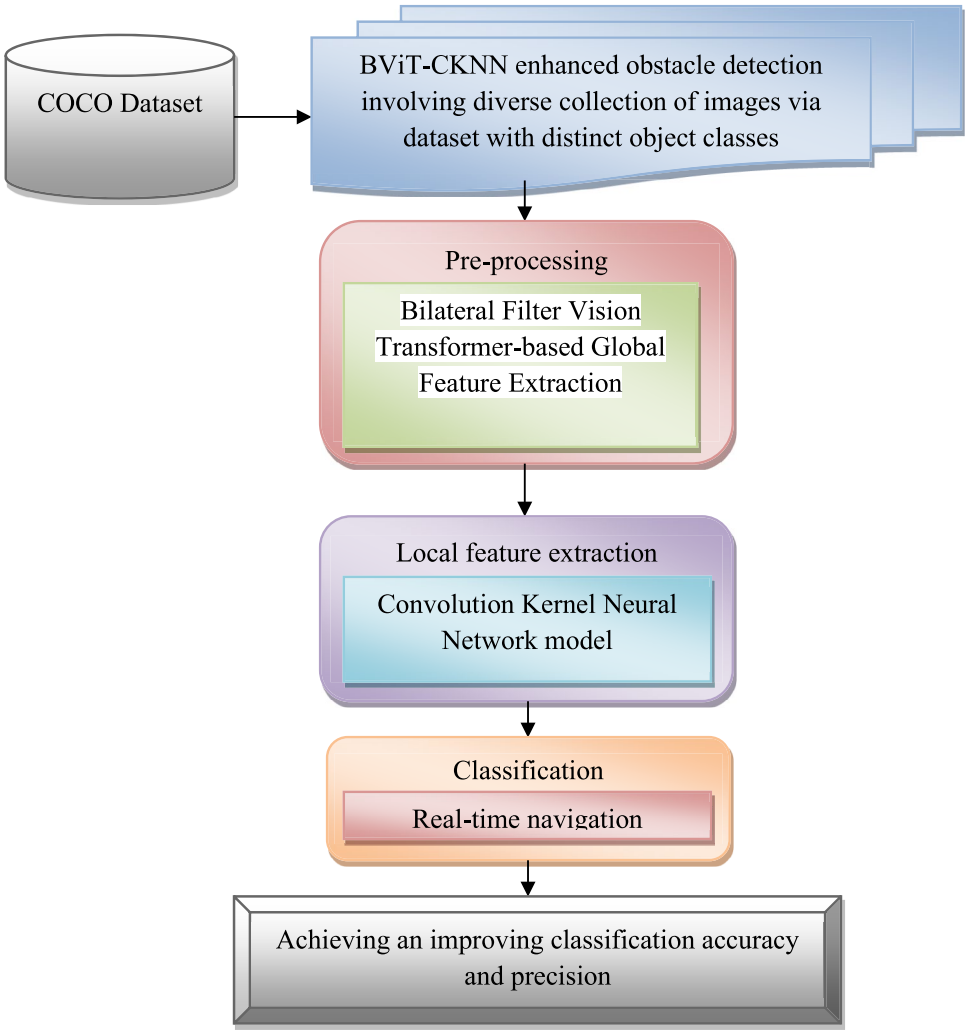

Fig. 1 Architecture diagram of the proposed Bilateral Vision Transformer and Convolution Kernel Neural Network (BViT-CKNN) method

Table 2 Categories in the COCO dataset

S. No	Categories	ID
1	Person ID	Id – 1
2	Vehicle	Id – 2 to 9
3	Outdoor	Id – 10 to 15
4	Animal	Id – 16 to 25
5	Accessory	Id – 26 to 33
6	Sports	Id – 34 to 43
7	Kitchen	Id – 44 to 51
8	Food	Id – 52 to 61
9	Furniture	Id – 62 to 70
10	Electronic	Id – 71 to 77
11	Appliance	Id – 78–82
12	Indoor	Id – 84–90

classes to their respective identification ranges. The Person category (ID 1) includes human figures critical for collision avoidance in crowded areas. The Vehicle category (IDs 2–9) includes various modes of transportation- bicycles, cars, and buses—essential for outdoor mobility assistance. Outdoor (IDs 10–15) comprises environmental elements such as trees and benches, while the Animal category (IDs 16–25) enhances the system's capability to detect dynamic, unpredictable obstacles like pets.

Other significant categories include Accessory (IDs 26–33) and Sports (IDs 34–43), which add to the dataset's contextual complexity. The Kitchen (IDs 44–51) and Food (IDs 52–61) classes support indoor navigation and scene recognition applications. Categories like Furniture (IDs 62–70), Electronic devices (IDs 71–77), and Appliances (IDs 78–82) reflect indoor object interactions, whereas Indoor (IDs 84–90) classes encompass general in-home items and spaces.

By training and evaluating the model on such a diverse and context-rich dataset, the proposed framework is expected to generalize well across a wide range of real-world scenarios, ensuring its practical applicability and reliability in assisting visually impaired users during navigation and obstacle avoidance.

4.3 Bilateral filter vision transformer-based global feature extraction

Preprocessing plays a vital role in obstacle detection for visually impaired individuals by enhancing image quality and reducing noise, which improves the accuracy of subsequent tasks like object detection and tracking. The proposed method employs a bilateral filter before processing with a Vision Transformer (ViT). This filter smooths the image while preserving important edges, unlike traditional filters that may blur them. By integrating bilateral edge-preserving smoothing with ViT, the system enhances image clarity, helping the model focus on relevant features better. The bilateral filter replaces each pixel's intensity with a weighted average of neighbouring pixels, where the weights are based on both spatial closeness and intensity differences—thereby ensuring edge preservation and effective noise reduction to support improved object detection performance.

Let 'SI' represents the Sample Training Image, and then the Bilateral Filter output image '$S^{filtered}$' is mathematically derived as given below.

$$\begin{aligned} SF &= S^{filtered}(p_i) \\ &= \frac{1}{Norm} \sum_{q_i \in \alpha} SI(p_{i+1})\, RK(||SI(p_{i+1}) - SI(p_i)||) \\ &\quad SK(||p_{i+1} - p_i||) \end{aligned} \tag{1}$$

$$\begin{aligned} Norm = \sum_{q_i \in \alpha} & RK(||SI(p_i) - SI(p_{i+1})||) \\ & SK(||p_i - p_{i+1}||) \end{aligned} \tag{2}$$

From the above Eq. (1), (2) sample filtered training image is denoted as 'SF', and weight normalization results are represented as '$Norm$'.The pixel intensity values in the sample training image are restored with a weighted mean of intensity values of the neighbouring pixel. Moreover, 'SI' and '$S^{filtered}$' represent the Sample Training Image and filtered image, respectively, based on the window centre 'α', 'p_i' indicates the coordinates of the current pixel to be filtered, and 'p_{i+1}' represented as a neighbouring pixel. Moreover, the bilateral filter function in our work utilizes spatial kernel 'SK' (i.e. smoothing differences in pixel spatial coordinates) and range kernel 'RK' (i.e. smoothing differences in pixel intensity) intensity values to preserve sharp edges of corresponding Sample Training Images based on different classes (i.e. 80 classes). Following this, the weight normalization results are arrived at using '$Norm$' in Eq. (2).

The proposed Vision-aided Transformer (ViT) in the ViT-NN method specifically consists of two paramount elements: a feature extractor and a classifier. Our work uses preprocessed or filtered samples for further processing (i.e., feature extraction and classification). Important features from filtered images were discovered. On the other hand, the task of the classifier is to split the sample-filtered training image into distinct classes. Moreover, in the proposed method, both global context understanding using a Bilateral Filter Vision Transformer and fine-grained local feature extraction using CNN are designed for feature extraction.

Transformer encoder layers employed with feature extractor. It comprises multi-head self-attention by position-wise feed-forward network. The proposed technique concentrates on numerous portions of sample-filtered training images and ascertains associations among them via self-attention. At input, every layer of the sequence obtains non-linear transformation.

Assume all patches are tokens with a sample-filtered training image 'SF" sequence. At first, the sample-filtered training image is separated into fixed-size patches. Every patch is converted to a vector. The sample filtered training image'SF' sequence is split into non-overlapping patches of fixed size '$16 * 16 pixels$' and converted into a vector as given below.

$$SF = \begin{bmatrix} SF_1 \\ SF_2 \\ SF_3 \\ \cdots \\ SF_{N-1} \\ SF_N \end{bmatrix} = \begin{pmatrix} SF_1 \\ SF_2 \\ SF_3 \\ \cdots \\ SF_{N-1} \\ SF_N \end{pmatrix} \tag{3}$$

From the above Eq. (3) sample filtered training image '$SF_i = \{SF_1, SF_2, \ldots, SF_{N-1}, SF_N\}$' sequence or image, dimensions '$h * w * C$' is split to 'N' non-overlapping patches of size '$P * P$'. '$N = \frac{(h*w)}{P^2}$' employed to estimate patches. With the converted vector according to the above (3), each patch is then flattened into a one-dimensional vector by employing the Learnable Projection Matrix as given below.

$$LM \in \mathbb{R}^{(P^2 * C) * D} \tag{4}$$

From the above Eq. (4), 'LM' is indicated as a linear projection matrix, '$\in \mathbb{R}^{(P^2)}$' is denoted as the input vector, 'D' denotes the dimensionality of embedding result, 'P^2 ' is represented as patch sizes and 'C' specifies classes. The flattened patch into a one-dimensional vector result is obtained using Eq. (4). The above-flattened vectors are then passed via linear projection matrix ('LM') to generate patch embedding results and are mathematically represented as given below.

$$PE = [SF_1 LM, SF_2 LM, \ldots, SF_N LM] + LM_{Pos} \tag{5}$$

From the above Eq. (5), the patch embedding results 'PE' are arrived at based on the linear projection matrix ('LM')

results of the first sample filtered training image 'SF_1LM', linear projection matrix ('LM') results of second sample filtered training image 'SF_2LM' and so on in addition to the positional encoding results of the linear projection matrix 'LM_{Pos}' respectively.

The proposed model learns spatial correlations among patches. With subsequent patch embedding, positional encoding of every token is included. Encoding results of the linear projection matrix '$LM_{Pos} \in \mathbb{R}^{N*D}$' also preserves spatial information that is said to be lost during the process of flattening. Lastly, important features are selected by patch embeddings and positional encodings. These broadcast to transformer encoder layers.

Figure 2 illustrates the preprocessing stage of the proposed BViT-CKNN obstacle detection framework, emphasizing the role of bilateral filtering in enhancing image quality before feature extraction. The process begins with an input image from the COCO dataset containing real-world objects and scenes with complex backgrounds. This image—the original image—is then passed through a bilateral filter function, a non-linear, edge-preserving, and noise-reducing smoothing technique.

Unlike traditional filters that often blur edges, the bilateral filter effectively smooths homogeneous regions while preserving crucial edge details. This capability is significant for obstacle detection systems targeted at visually impaired individuals, where precise boundary recognition and edge clarity are critical for accurate object localization. The output is a filtered image that maintains strong object outlines and suppresses unnecessary background noise, creating an optimal input for the subsequent feature extraction modules. This enhanced visual input significantly improves the performance of deep learning models by allowing them to focus on meaningful structures within the image, ultimately contributing to more accurate and reliable obstacle detection.

4.4 Convolution Kernel neural network-based local feature extraction

Convolution Kernel Neural Network (CKNN) as a core local feature extractor. It isn't a standard term in deep learning and denotes to core functionality of CNNs, where convolutional kernels (or filters) act as local feature extractors. CNNs leverage these kernels to identify patterns within local regions of input data, such as images, and this process is fundamental to their ability to learn hierarchical feature. The CKNN is not a standard or extensively recognized in deep learning literature. It appears to be a misnomer or a localized term for a concept that can be described using established terminology.A standard CNN employs convolutional layers with a set of kernels to extract the features. Each kernel is considered to identify the specific patterns or features within a local accessible field. The CKNN as a core local feature extractor aligns directly with the fundamental purpose of convolutional layers in CNNs.

The specific calculation process of Convolution Kernel-based Local Feature Extraction is given below. For each parallel convolution layer, an '$N-gram$' convolution between patch embedding results and different sized convolution kernels '$CK = (CK_1, CK_2, \ldots, CK_D)$' is performed. '$CK_i \in \mathbb{R}^{K*D}$' with '$K$' representing the kernel size and 'D' denoting the dimensionality of embedding. Figure 3 shows the structure of Convolution Kernel-based Local Feature Extraction.

As shown in the above figure, in the Convolution Kernel-based Local Feature Extraction model implementation. Various extractions '$N-gram$' feature classes are included in

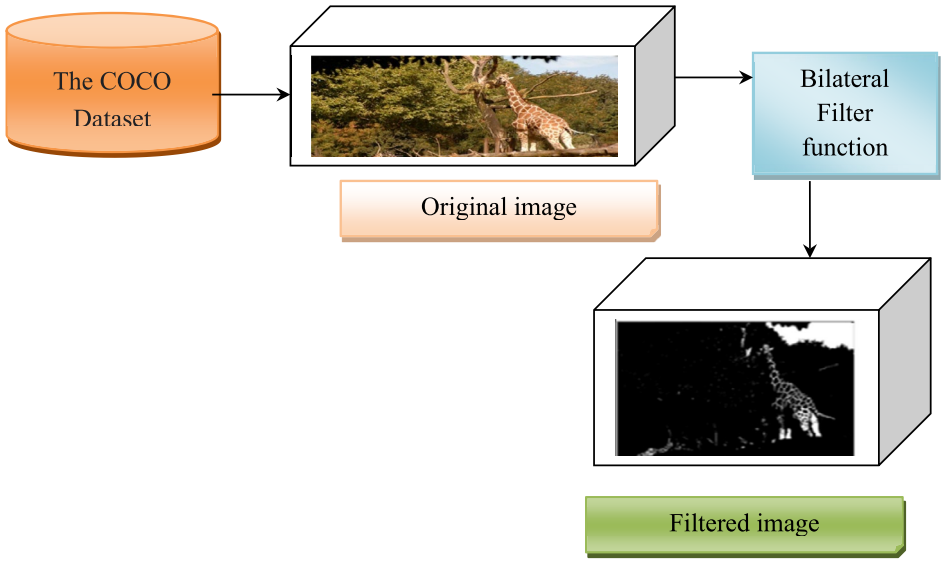

Fig. 2 Structure of Bilateral Filter based preprocessing model

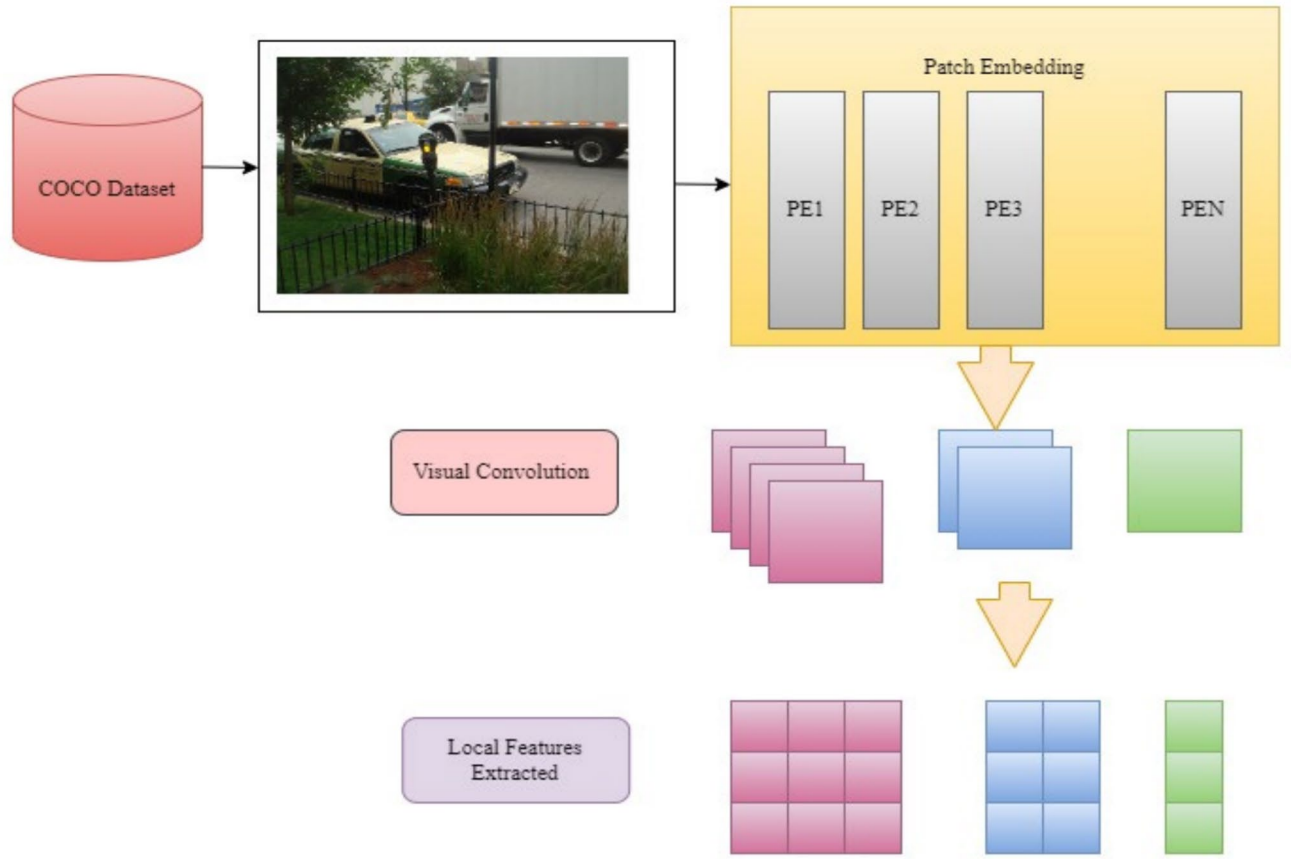

Fig. 3 Structure of Convolution Kernel-based Local Feature Extraction model

several convolutional kernels of divergent sizes. For local features, an image combination of different classes is captured via distinct kernel sizes. Visual feature map matrix 'VFM_l' is mathematically formulated as given below.

$$VFM_l = PE \circledast CK_i \tag{6}$$

From the above Eq. (6),'⊛' represents the convolution operation of 'CK_i' on patch embedding results 'PE'. 'VFM_l 'is denoted below.

$$VFM_i = f\left(CK_i * PE + f\right) \tag{7}$$

From Eq. (7), 'VFM_i'is indicated as local fine-grained features. Employing a non-linear activation function 'f', the Convolution Kernel maps the original patch embedding results to a visual feature map highlighting specific features or patterns (i.e. edges). In this manner, the output matrices of the convolution layers '$VFM_1, VFM_2, \ldots VFM_l$' are obtained. These output matrices highlight specific features or patterns (i.e., edges) and better capture intrinsic obstacles, thereby improving overall performance.

4.5 Transformer encoder-based global feature extraction

The transformers is factor of object detection systems where IoU is crucial, the IoUestimate itself is not an internal operation of the Transformer encoder during its feature extraction process. The encoder focuses on learning inclusive feature representations, which are then analyzed by subsequent modules (like prediction heads) to leverage IoU for tasks such as bounding box regression and evaluation.

From above Fig. 4, the Transformer encoder –based Global Feature Extraction is described. Transformer encoder layers execute different patch sequence 'PE'. Two chief elements are used in each encoder layer. Local and global figures obtained via self-attention. The self-attention mechanism permits the model to assess the significance of each patch in proportion to all others, obtaining both local and global contextual information. The Transformer encoder's main role in feature extraction is to capture global dependencies and relative information within input sequence (e.g., image patches in vision transformers or tokens in natural language processing). The IoU is a standard metric for evaluating the performance of object detection models,

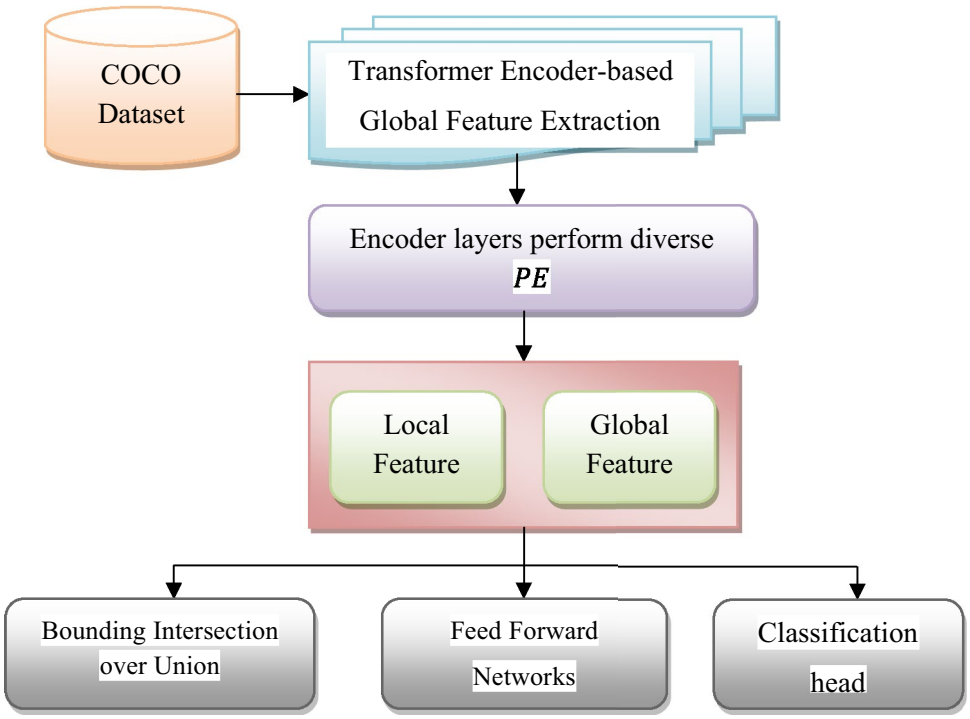

Fig. 4 Transformer Encoder-based Global Feature Extraction

where a high IoU between predicted and ground-truth boxes indicates accurate localization.

4.5.1 Bounding intersection over union-based multi-head self-attention

In the context of enhanced obstacle detection and computer vision, utilizing multi-head self-attention with Intersection over Union (IoU) for measuring bounding box exploits the potentiality of self-attention mechanisms to boost object localization and detection accuracy, frequently by fine-tuning bounding box predictions on the basis of contextual information and optimizing IoU for improving object detection performance. A bounding box, being a shape, highlights an object in an image.

The bounding box possesses several features, including bounding box height 'BB_h', bounding box width 'BB_w ', and bounding box class 'C'. Then, bounding box centre coordinates '(p_{center}, q_{center})' by 'BB_w' and 'BB_h' are obtained as given below.

$$p_{center} = PE\left[\frac{BBC_p}{SF_w}\right] \tag{8}$$

$$q_{center} = PE\left[\frac{BBC_q}{SF_h}\right] \tag{9}$$

From the above Eqs. (8) and (9), centre coordinates of the bounding box '(p_{center}, q_{center})' results are arrived at based on the bounding box center pixel width 'BBC_p', bounding box center pixel height 'BBC_q', sample filtered training image width 'SF_w', sample filtered training image height 'SF_h'.

$$BB_w = PE\left[\frac{BBC_w}{S_w}\right] \tag{10}$$

$$BB_h = PE\left[\frac{BBC_h}{S_h}\right] \tag{11}$$

Similarly, from the above Eqs. (10) and (11), the bounding box coordinate width 'BB_w' and bounding box coordinate height 'BB_h' results are arrived at using the bounding box centre width 'BBC_w', bounding box centre height 'BBC_h ' respectively.

$$Res\,(SF_i) = [BB_w, BB_h, (p_{center}, q_{center})] \tag{12}$$

$$\begin{aligned} IoU\,[Res\,(SF_i)] &= \frac{Areaofoverlap}{Areaofunion} \\ &= \frac{|Res\,(SF_p) \cap Res\,(SF_q)|}{|Res\,(SF_p) \cup Res\,(SF_q)|} \end{aligned} \tag{13}$$

Finally, from the above equation, the Intersection over Union results '$IoU\,[Res\,(SF_i)]$' are arrived at based on the degree of overlap between bounding boxes. An '$IoU\,[Res\,(SF_i)]$ ' score of '1' indicates perfect overlap 'X', denoting predicted and ground truth bounding regions being identical, hence forming the input to the transformer encoder. On the other hand, the'$IoU\,[Res\,(SF_i)]$' score of '0'indicates no

overlap, denoting the predicted and ground truth bounding regions found to be entirely disjoint, hence not considered for further processing.

Assume input 'X' on layer 'l', the self-attention function measures attention scores via three linear projections 'Q_l', 'K_l', and 'V_l' as given below.

$$Q_l = XW_Q^l \tag{14}$$

$$K_l = XW_K^l \tag{15}$$

$$V_l = XW_V^l \tag{16}$$

From the above Eqs. (14), (15) and (16), 'W_Q^l', 'W_K^l', and 'W_V^l' denote the learning weight matrices for the corresponding resultant sample filtered images at layer 'l'. Following the results obtained above, the attention scores are evaluated with the aid of scaled dot-product attention, as given below.

$$Attention\left(Q_l, K_l, V_l\right) = Softmax\left(\frac{Q_k K_l^T}{\sqrt{d}}\right) V_l \tag{17}$$

From the above equation results (17), attention scores of three linear projections '$Attention\left(Q_l, Kl, V_l\right)$' were measured. It provides multiple attention head outputs, which are represented below.

$$Multi - head\left(X_l\right) = \left[Head_1, Head_2, \ldots, Head_H\right] * W_l^O \tag{18}$$

From the above Eq. (18), multi-head attention results '$Multi - head\left(X_l\right)$' are arrived at based on the 'H' number of attention heads '$Head_H$' and learning weight matrix 'W_l^O', respectively.

4.5.1.1 Feed forward networks Multiple attention heads and position-wise feed-forward networks are used for patch embedding.

$$FFN\left(X_l\right) = GELU\left(X_l W_1^l + b_1^l\right) * W_2^l + b_2^l \tag{19}$$

From the above Eq. (19), the feed-forward network'$FFN\left(X_l\right)$' results of the 'lth' perfect overlap 'X' results are arrived at based on the weight matrices 'W_1^l', 'W_2^l' and bias matrices 'b_1^l', 'b_2^l' activated via the '$GELU$' function. Residual connection and layer normalization are measuredby feature vector sequence as given below.

$$X_{l+1} = LayerNorm\left(FFN\left(Multi - head\left(X_l\right)\right) + X_l\right) \tag{20}$$

From the above Eq. (20), 'X_{l+1}' denotes the '$l+1$' perfect overlap 'X' results.

4.5.1.2 Classification head The classifier is carried out to predict sample training input image class labels. It is used for generated patch embedding. The classified results are obtained using a softmax activation function.

$$SO = Softmax\left(X_L^1 W_C + b_C\right) \tag{21}$$

From the above Eq. (21), the predicted class probability result 'SO' is arrived at based on the total number of encoder layers 'L', 'W_C' and 'b_C' forming the weight matrix for classification and bias for classification, respectively. The flow diagram of the proposed BViT-CKNN is illustrated in Fig. 4.

The pseudo-code representation of enhanced obstacle detection for visually impaired people using a Bilateral Vision-aided Transformer Neural Network is given below.

Input: Dataset ' DS ', Sample Training Images ' $SI = \{S_1, S_2, \ldots, S_N\}$ ', Classes ' $C = \{C_1, C_2, \ldots, C_M\}$'

Output: Fine-grained feature extraction with precise and error-minimized enhanced obstacle detection

1: **Initialize** '$N = 25000$', '$M = 80$'
2: **Begin**
3: **For** each Dataset 'DS' with Sample Training Images 'S' and Classes 'C'
// Preprocessing using Bilateral Filter Vision Transformer
4: Estimate Bilateral Filter output image using equation (1)
5.Derived weight normalization results using equation (2)
6. Obtain preprocessed filtered images
7. Sample filtered image is separated into fixed-size patches
5: Convert each patch into vector according to equation (3)
6. **For** each patch Flattened into one dimensional vector using Learnable Projection Matrix in equation (4)
7: Formulate linear embedding and Positional Encoding results according to (5)
//Convolution Kernel-based Local Feature Extraction
8: Evaluate visual feature map matrix according to (6)
9: Obtain local fine-grained features according to (7)
10: Return local fine-grained features 'VFM_i'
//Transformer Encoder-based Global Feature Extraction
//Positional Encoding
11: Evaluate center coordinates of bounding box '(p_{center}, q_{center})', bounding box width 'BB_w' and height 'BB_h' according to (8), (9), (10) and (11)
13: Measure Intersection over Union results according to (13)
14: **If** '$IoU\ [Res\ (SF_i)] = 1$'
15: **Then** go to step 21
16: **End if**
17: **If** '$IoU\ [Res\ (SF_i)] = 0$'
18: **Then** go to step 6
19: **End if**
20: **End for**
//Transformer Encoder
21: **For** each Dataset 'DS' with positional encoded results '$Res\ (SF_i)$' and Classes 'C'
//Bounding Intersection Over Union-based Multi-Head Self-Attention
22: Measure the attention scores for each patch via three linear projections ' Q_i ', ' K_i ', 'V_i'according to (14), (15) and (16)
23: Evaluate attention scores are evaluated with the aid of scaled dot-product attention according to (17)
24: Evaluate multiple attention heads according to (18)
//Feed Forward Networks
25: Evaluate position-wise feed-forward network to each patch embedding independently according to (19)
26: Evaluate global feature extractor results according to (20)
//Classification Head
27: Evaluate classified results via softmax activation function according to (21)
28: **Return** sample output 'SO'
29: **End for**
30: **End**

Algorithm 1 Bilateral Vision-aided Transformer Neural Network for enhanced obstacle detection with visually impaired people

As given in the above algorithm, with the objective of enhancing the precision and accuracy involved in enhanced obstacle detection with visually impaired people in a computationally efficient manner, the overall process is split into three parts. First, the COCO (Common Objects in Context) dataset obtained from https://www.kaggle.com/code/armanasgharpoor1993/coco-dataset-tutorial-image-segmentation/notebook as input is subjected to preprocessing using Bilateral Filter Vision Transformer. Using this model, the raw input images in the corresponding objects are considered edge preserved and minimize noise via a non-linear function, reducing the overall mean square error. Second, the obtained sample filtered training image is then flattened, and then Linear Embedding and Positional Encoding functions are applied where a dimensional image is converted into a single dimension via the flattening process and spatial information being lost during flattening is preserved via Positional Encoding.Following this Convolution, a Kernel-based Local Feature Extraction model is applied to the flattened and encoded images to extract fine-grained local features accurately. Finally, the Transformer Encoder-based Global Feature Extraction model is applied to the flattened and encoded images to extract global contextual images precisely. Here, using Bounding Intersection Over Union-based Multi-Head Self-Attention aids in considerably improving precision and recall factors. This, in turn, ensures that the classified output results are computationally efficient, hence paving enhanced object detection mechanisms for visually impaired people.

5 Case analysis and inferences

In this section, case analysis of enhanced obstacle detection using different classes of objects obtained from the COCO dataset is simulated by applying the Bilateral Vision Transformer and Convolution Kernel Neural Network (BViT-CKNN) method. Figure 5, given below, shows the sample classes of images, namely, vehicles, animals, food, furniture, and electronics.

Five different classes of objects (i.e. samples) are used to perform the simulation, as shown in the above figure. With a dataset consisting of 80 distinct object classes, including common objects like vehicles, outdoor, indoor, animal, food and more, each object class is associated with a unique ID. In our simulation scenario, vehicle ID (2 to 9), animal ID (16 to 25), food ID (52 to 61), furniture ID (62 to 70) and electronic ID (71 to 77) are used for enhanced obstacle detection. The bilateral filter function, by employing both spatial kernel and range kernel intensity values, preserves sharp edges for different classes (i.e. 80 classes). This, in turn, aids in filtering the irrelevant portion of the image and retaining the essential portion of the image for further processing, therefore reducing overall mean absolute error. Figure 6 illustrates the output of preprocessed results for the different classes (a) vehicle, (b) animal, (c) food, (d) furniture, and (e) electronics, respectively.

Five different classes of objects are provided as input in the above figure, and they have varying levels of complexity. The above preprocessed results are provided as input for obtaining fine-grained local feature extraction employing

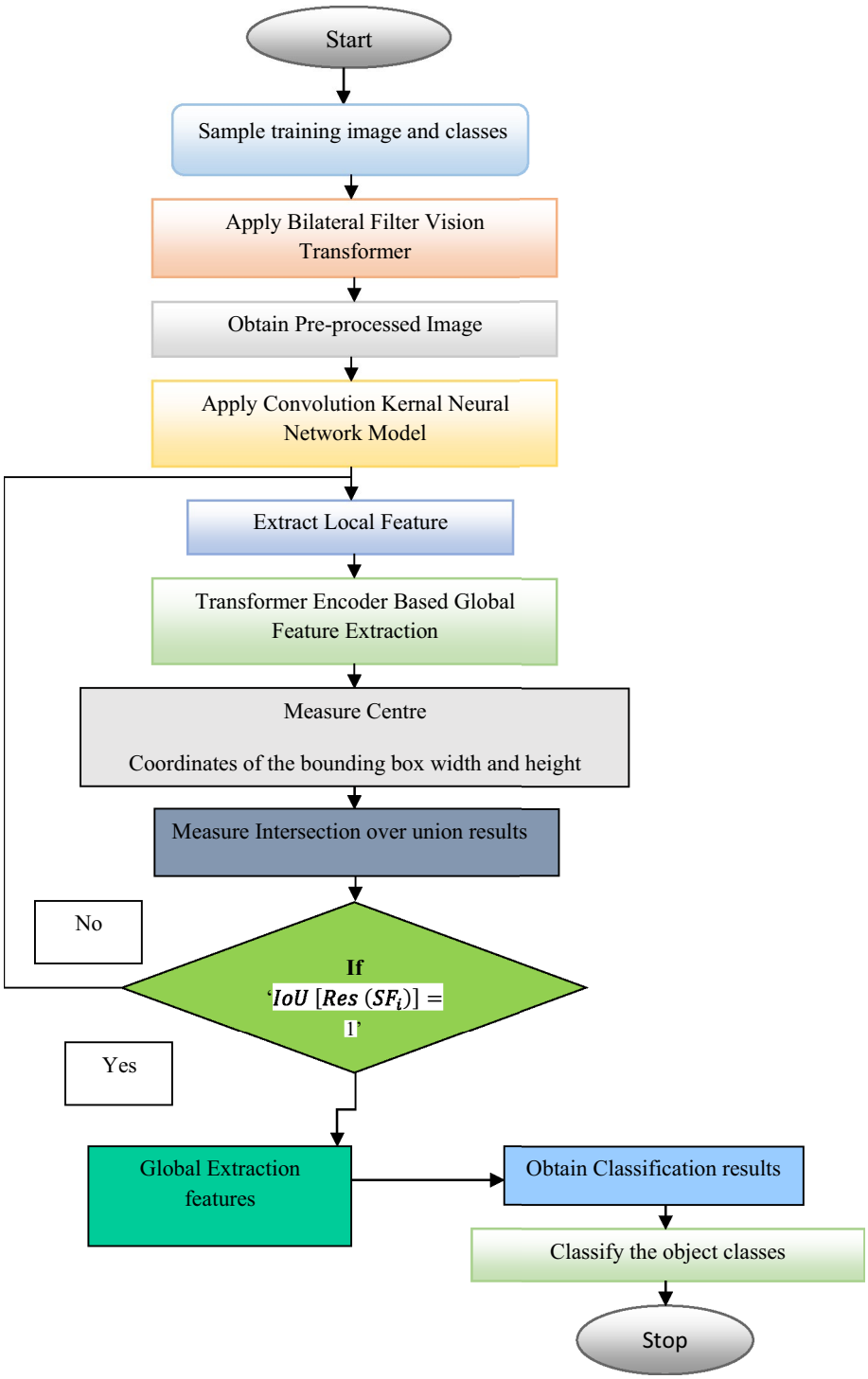

Fig. 5 Flow chart diagram of proposed BViT-CKNN

Fig. 6 Input sample images collected from the COCO dataset with five different classes (i.e. vehicle, animal, food, furniture and electronics)

the Convolution Kernel Neural Network model. The results are given in Fig. 7.

Figures 7 (a) and (b) given above show the local and global context feature extracted results when applied with vehicle preprocessed images, (c), (d) representing the local and global context feature extracted results of animal preprocessed images and so on for three more different classes of objects. Here, first, the preprocessed sample classes of objects with the objective of improving the model's effectivenessvisual feature map matrix were employed, therefore paving the way for accurate and precise obstacle detection results. Here, applying both the local and global feature extracted results to classification, in turn, aids in detecting different obstacles, therefore aiding visually impaired individuals precisely. Hence, summarizing both local and global image features accurately and precisely describes enhanced obstacles with varying levels of complexity. These varying complexities are handled using Bounding Intersection Over Union-based Multi-Head Self-Attention. With this, the overall precision and recall rate were found to be improved by addressing varying levels of complexities via multi-head self-attention of the transformer encoder layer.

6 Experimental and setup

This section discusses the parametric analysis for proposed enhanced obstacle detection for handling visually impaired people with the aid of Bilateral Vision Transformer and Convolution Kernel Neural Network (BViT-CKNN). Simulations are performed in MATLAB-based graphical programming environment for modeling, simulating and analyzing multi-domain dynamical systems on a on a computer Intel(R) Core (TM) i7-6700HQ CPU@2.60 GHz with a RAM of 32 GB running Windows. The initial split provides training (83 K), validation (41 K) and test (41 K) sets.

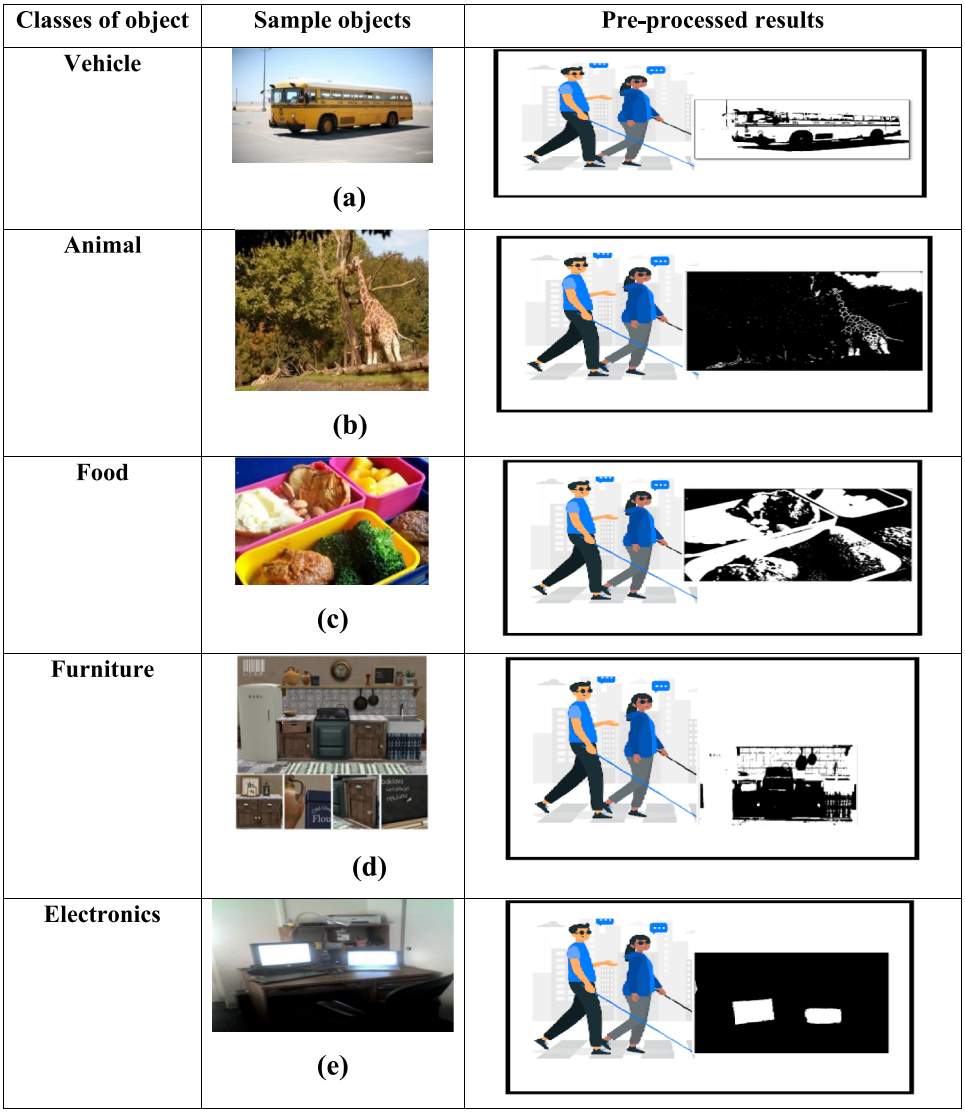

Fig. 7 five different classes of objects (i.e. vehicle, animal, food, furniture and electronics) of input images applied to Bilateral Filter preprocessing for obtaining the preprocessed results

The model Inference Time (ms/image) depends on deep learning method, hardware setup is GPU/CPU, embedded device Vs desktop, model size is 224 × 224, Time is describe the 15–30 ms and Model Size (MB): 5 to 15 MB.Fair comparison analysis is made using the four methods, BViT-CKNN, modified YOLO v5 neural network[1],SSDLite MobileNetV2 [2], and State-of-art (SOTA) work, namely BLIP, where the same objects possessing different classes of images are used for validating the performance metrics. Five different classes, namely, vehicle, animal, food, furniture, and electronic, are considered for the performance of simulations.

6.1 Precision and recall

It is presented to enhance obstacle detection for visually impaired people.

$$Pre = \frac{TP}{TP + FP} \tag{22}$$

$$Rec = \frac{TP}{TP + FN} \tag{23}$$

From the above Eqs. (22) and (23) 'Pre' is precision, 'Rec' is recall, 'TP' is true positive (i.e., vehicle sample objects detected as vehicle objects), false positive rate'FP'(i.e. vehicle sample objects detected as animal object) false negative rate'FN' (i.e. animal sample objects detected as vehicle object) respectively. Table 3 given below lists the precision and recall analysis employing BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP.

Figures 8, 9 given above shows the precision results with respect to 150,000 different sample images with respect to five distinct classes obtained from the Common Objects in Context dataset for the proposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP. From the above figure with simulation performed for 15,000 images, the true positive rate using the three methods were observed to be 13,700, 13,000, 12,500, and 11,000, whereas the false positive rate using the three methods was found to be 300, 1000, 1500 and 3000 respectively. Precision was observed to be 97%, 92%, 89%, 78% or BViT-CKNN, [1, 2] and BLIP. Better results were provided for precision using the proposed BViT-CKNN method.

Figure 10 given above shows the recall rate results with respect to 150,000 different object categories obtained from the COCO dataset for the three different methods, BViT-CKNN, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP respectively. From the above figure with simulation performed for 15,000 images, the true positive rate using the three methods was observed to be 13,700, 13,000, 12,500 and 11,000, whereas the false negative rate using the three methods was found to be 200, 300, 400 and 600 respectively. With this, the overall recall rate when applied with the three methods was found to be 98%, 97%, 96% and 94%, respectively. This, in turn, confirms comparative better results for recall rate using the proposed BViT-CKNN method.

Contrary to conventional methods, from the above two results of precision and recall, comparative results showed betterment using the proposed BViT-CKNN method. It hasbeen improved byseparate local feature extraction and global feature extraction. By applying the Convolution Kernel-based Local Feature Extraction model, flattened and encoded images were used to extract fine-grained local features. To enhance object localization and detection, Bounding Intersection over Union was employed.Combining these two ensured enhanced obstacle detection even in the presence of five different classes (i.e. vehicle, animal, food,and furniture, electronic for simulation). With this, the overall precision using the proposed BViT-CKNN method was Said to be improved by 6% upon comparison to [1], 18% upon comparison to [2] and 24% upon comparison to BLIP.

Table 3 Tabulation of precision and recall using BViT-CKNN, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP

Samples	Precision				Recall			
	BViT-CKNN	modified YOLO v5 neural network [1]	SSDLite Mobile-NetV2 [2]	BLIP	BViT-CKNN	modified YOLO v5 neural network [1]	SSDLite Mobile-NetV2 [2]	BLIP
15,000	0.97	0.92	0.89	0.78	0.98	0.97	0.96	0.94
30,000	0.95	0.9	0.8	0.76	0.96	0.86	0.81	0.78
45,000	0.93	0.88	0.78	0.73	0.93	0.83	0.78	0.75
60,000	0.9	0.85	0.75	0.71	0.91	0.81	0.76	0.7
75,000	0.88	0.82	0.72	0.69	0.89	0.79	0.74	0.68
90,000	0.91	0.86	0.76	0.7	0.87	0.77	0.72	0.68
105,000	0.93	0.88	0.78	0.72	0.88	0.78	0.73	0.69
120,000	0.95	0.9	0.8	0.74	0.9	0.8	0.75	0.72
135,000	0.95	0.9	0.8	0.75	0.92	0.82	0.77	0.74
150,000	0.97	0.92	0.82	0.78	0.95	0.85	0.8	0.76

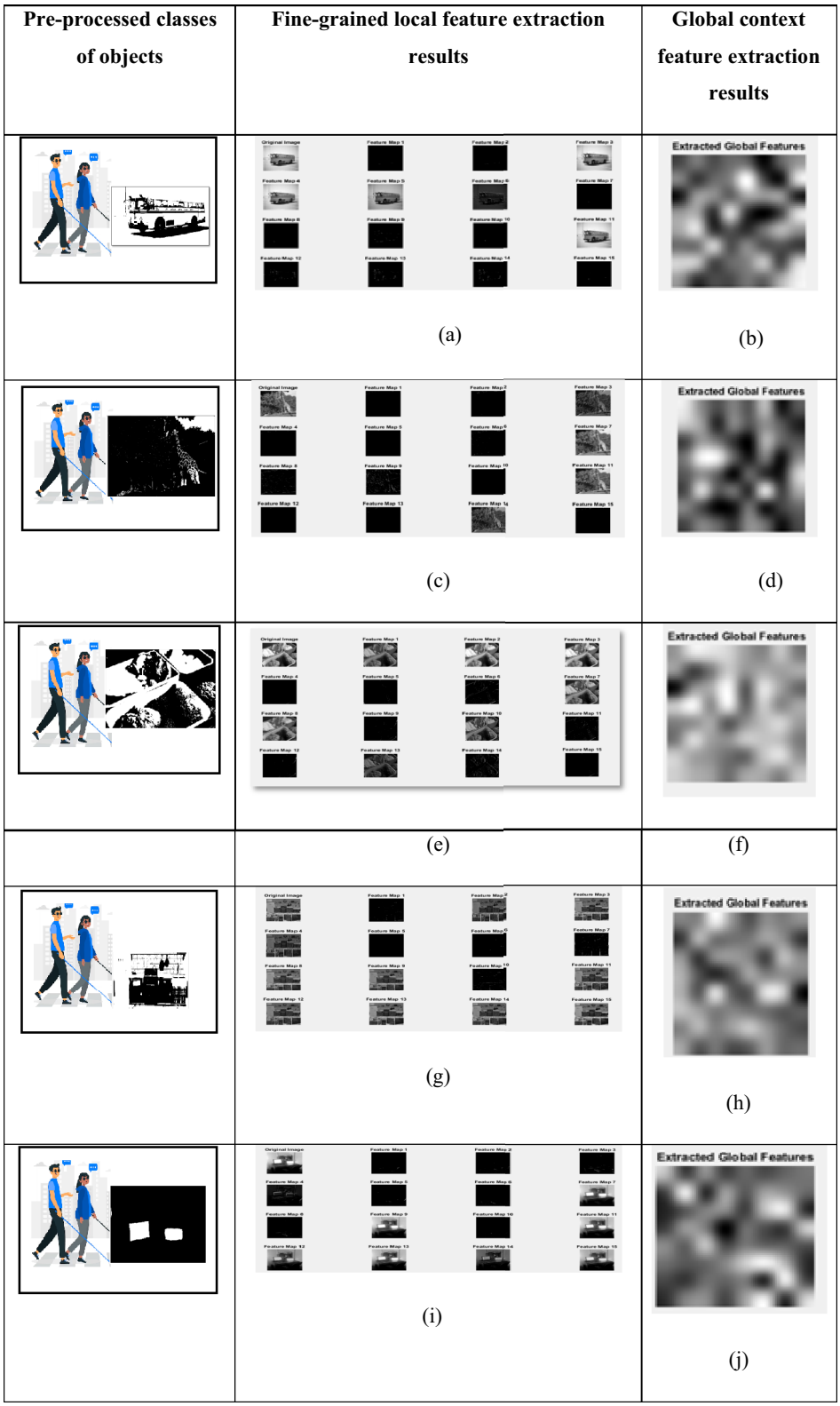

Fig. 8 Preprocessed results as input by using Convolution Kernel-based Local Feature Extraction and Transformer Encoder-based Global Feature Extraction to extract fine-grained local and global context feature extracted results for enhanced object detection

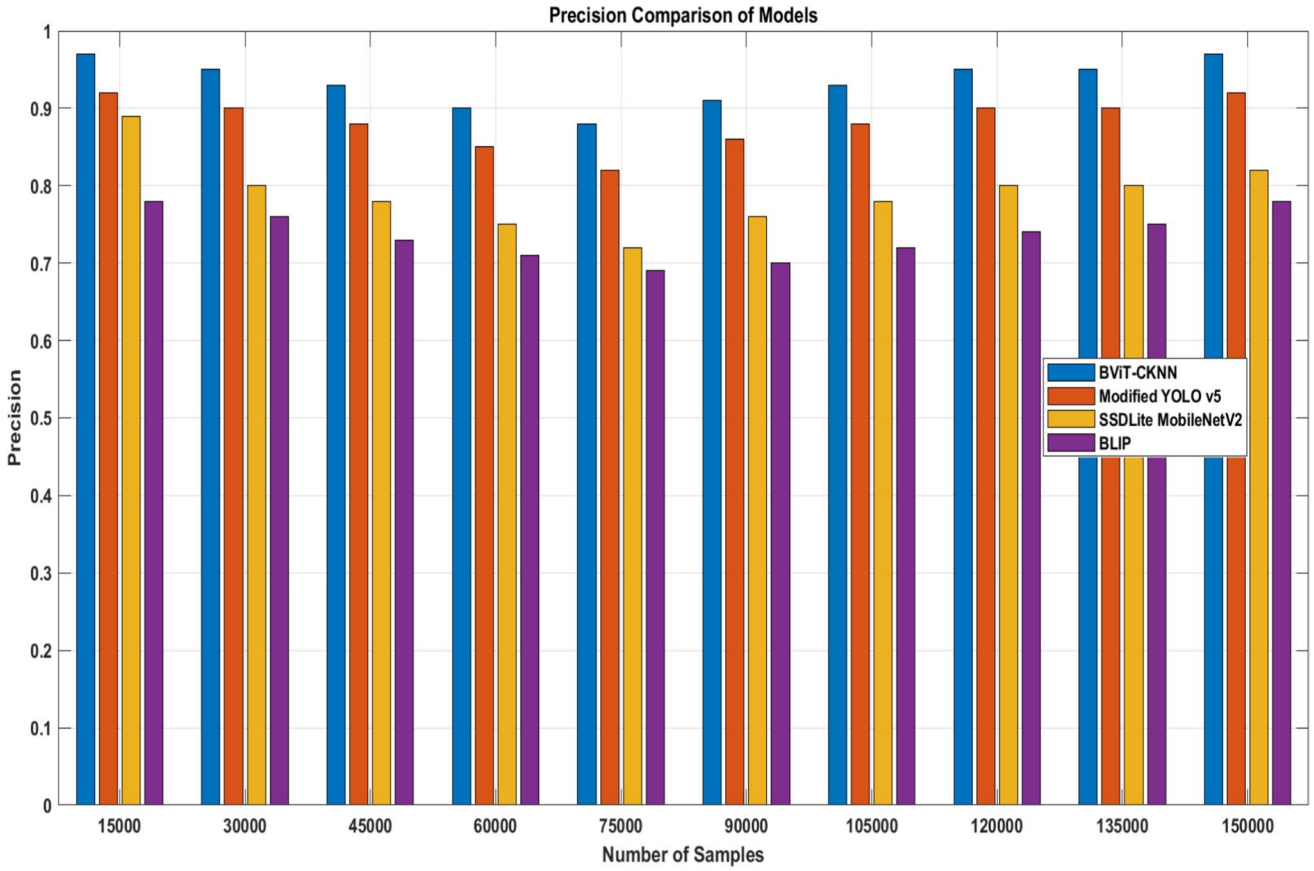

Fig. 9 Graphical representation of precision with number of samples forproposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP

BViT-CKNN of recall was found to be increased by 11%, 18%and 24% than [1, 2] and BLIP.

6.2 F1-score

Second, in this section, the F1-score analysis is made. The metric for Evaluation of Translation with Explicit Ordering (METEOR) score is used for the evaluation of machine-translation output. METEOR metric is considered for both harmonic means of precision and recall.In statistical analysis of binary classification (i.e., detection of enhanced obstacles for visually impaired people), the F1-score or F-measure is a measure of predictive performance. The F1-score is mathematically formulated as given below.

$$F1 - score = 2 * \frac{Pre * Rec}{Pre + Rec} \tag{24}$$

Table 4, the F1-score using proposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP, respectively.

Figure 11, given above, illustrates the F1-score with respect to 150,000 different sample images provided as input. From the above figurative representation, the F1-score observed using the three methods was found to be neither directly proportionate nor inversely proportionate to sample images considered as input. From the above simulations performed, the BViT-CKNN of the highest F-score value has0.97, whereas [1, 2], and BLIP were found to be 0.94, 0.92 and 0.85. This simulation result corroborates the objective of the highest F1-score values using the proposed BviT-CKNN method when compared to [1, 2] and BLIP.

6.3 Mean Absolute Error (MAE)

Depth estimation outlines the procedure of estimating distances from sensor data with varying levels of complexity, obviously in a two-dimensional array of depth range data. The varying levels of complexity here include multiple objects, occlusion and different backgrounds. MAEdeterminesstandardtotaldissimilarityamongforecastedand actualranges in image processing and depth estimation. The Mean Absolute Error is mathematically formulated as given below.

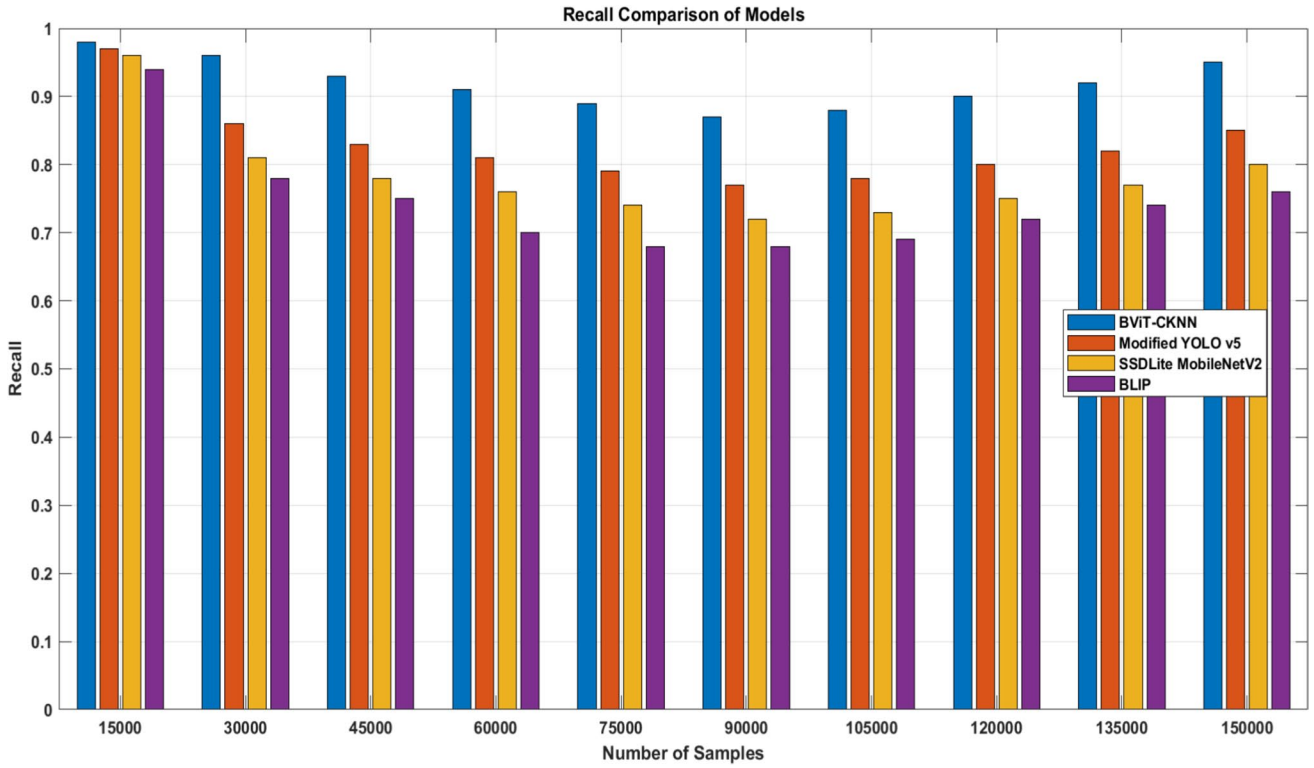

Fig. 10 Performance analysis of recall with number of samples for proposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP

Table 4 Tabulation of F1-score using BViT-CKNN, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP

Samples	F1-score			
	BViT-CKNN	modified YOLO v5 neural network [1]	SSDLite MobileNetV2 [2]	BLIP
15,000	0.974974	0.944339	0.923676	0.852558
30,000	0.954974	0.879545	0.804969	0.76987
45,000	0.93	0.854269	0.78	0.739865
60,000	0.904972	0.829518	0.754967	0.704965
75,000	0.884972	0.80472	0.729863	0.684964
90,000	0.889551	0.812515	0.739459	0.689855
105,000	0.904309	0.826988	0.754172	0.704681
120,000	0.924324	0.847059	0.774194	0.729863
135,000	0.934759	0.85814	0.784713	0.744966
150,000	0.959896	0.883616	0.809877	0.76987

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(SI_i - SO_i) \tag{25}$$

From the above Eq. (25), the Mean Absolute Error 'MAE' is measured based on the actual sample input 'SI_i' for the '$i-th$' observation and classified sample output 'SO_i' for the '$i-th$' observation with respect to a total number of 'N' samples. Table 5, given below, shows the mean absolute error using proposed BviT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP, respectively. See Fig. 12.

MAE is shown in Fig. 11. Improving the sample size also causes a proportionate increase in MAE. However, simulations performed for ten different iterations show the error found to be minimized using the proposed BViT-CKNN method upon comparison to [1, 2] and BLIP. With 15,000 samples provided as input, the correct classified sample output using the proposed BViT-CKNN method was observed to be 14,750, whereas using [1, 2], and BLIP was found to be 14,600, 14,500 and 14,400, respectively. With this, the overall mean absolute error using the three methods was observed to be 1.66%, 2.66% [1], 3.33% [2] and 4% BLIP, respectively. The reason behind the minimization of MAE using the proposed BViT-CKNN method was the application of a Bilateral filter-based preprocessing model. Applying this preprocessing not only smoothens the image but also preserves the edges and reduces noise. MAE of BViT-CKNN decreased by 21%, 41%, and 48% over [1, 2], and BLIP.

6.4 Ablation study

Ablation studies are utilized in computer vision to develop obstacle detection for visually impaired people's lives. The ablation study is the concept of removing a certain part of

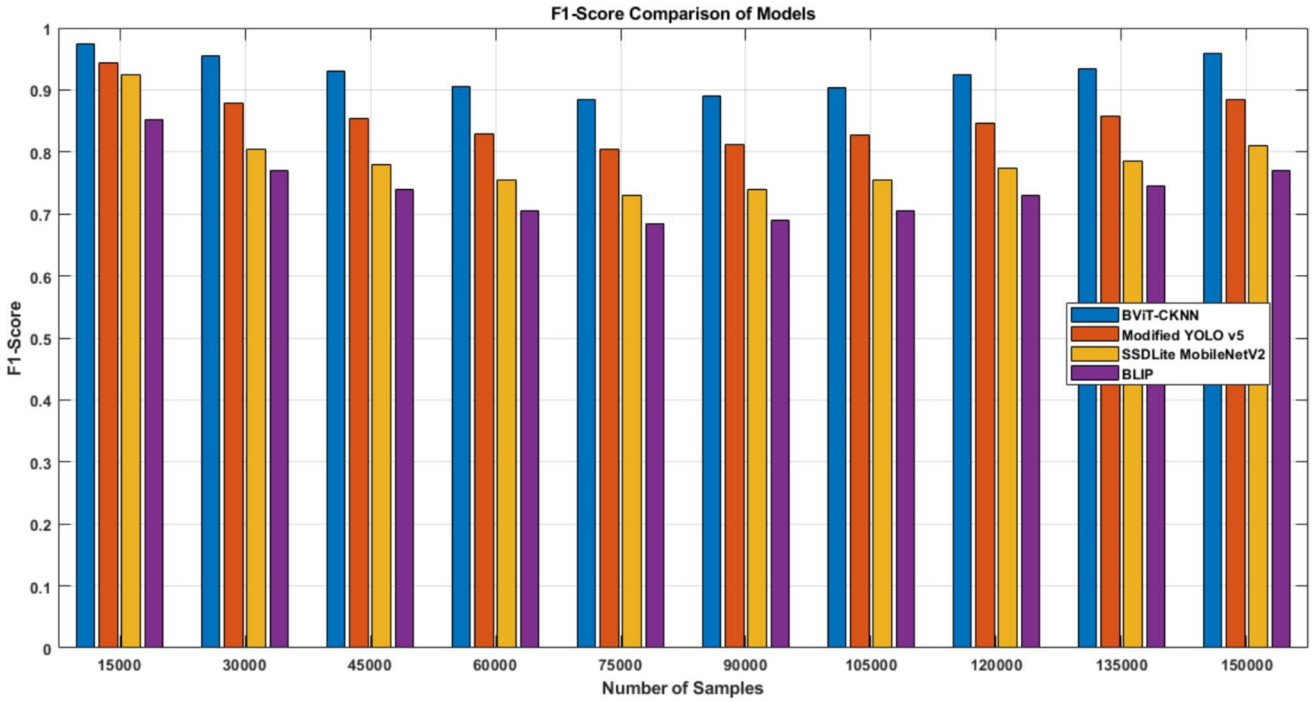

Fig. 11 Result of F1-score with number of samples for proposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP

Table 5 Tabulation of mean absolute errorusing BviT-CKNN, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP

Samples	Mean absolute error			
	BviT-CKNN	modified YOLO v5 neural network [1]	SSDLite MobileNetV2 [2]	BLIP
15,000	1.66	2.66	3.33	4
30,000	1.85	2.85	3.85	4.55
45,000	2.35	3	4.35	4.85
60,000	2.55	3.35	4.95	5.45
75,000	2.95	3.55	5.25	5.75
90,000	3.35	3.85	5.55	6.25
105,000	3.85	4.25	5.85	6.65
120,000	4.15	4.85	6	7
135,000	4.35	5.15	6.36	7.35
150,000	4.55	5.55	6.84	7.65

the network to get a better understanding of the network behaviour. In addition, an ablation study in the context of Vision Transformer (ViT) named deep learning is employed to determine each component's importance or contribution in a neural network model. It removes or"ablates"certain parts of the model, such as a specific layer or neuron, and observes the resulting impact on the method result. An ablation study is conducted, to examine the significance of each contribution involved in the proposed BViT-CKNN method. Table 6 given below lists the ablation study employing proposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2], and BLIP.

Table 6 shows the ablation study outcomes for four methods using the COCO dataset. The proposed BViT-CKNN method is divided into three major parts, namely preprocessing, local and global feature extraction, and classification. In an ablation study, portions of the samples are consistently removed to find significant images in the dataset. Initially, samples from the given database are taken as input. Preprocessing is performed with the Bilateral Filter Vision Transformer model to remove noise images in a dataset. After that, global and local features are extracted to classify the results. From the chosen extracted features, dissimilar object classes are detected in a precise manner. The results of the BViT-CKNN method are achieved with higher precision, recall, and F1-score by 0.93, 0.91, and 0.92, respectively, compared to existing methods.

6.5 Discussion

This study compares the proposedBViT-CKNN method with the existing [1] and [2], which are discussed with the COCO dataset based on various parameters, such as precision, recall, F1 score and Mean Absolute Error. The proposed BViT-CKNN method is evaluated with different performance metrics, namely, precision, recall, F1 score and Mean Absolute error with respect to different numbers of samples. The results confirm that the proposed BViT-CKNN method improved precision by 16%, recall by 18%, F-score by 0.97%, and Mean Absolute Error reduced by 37%when

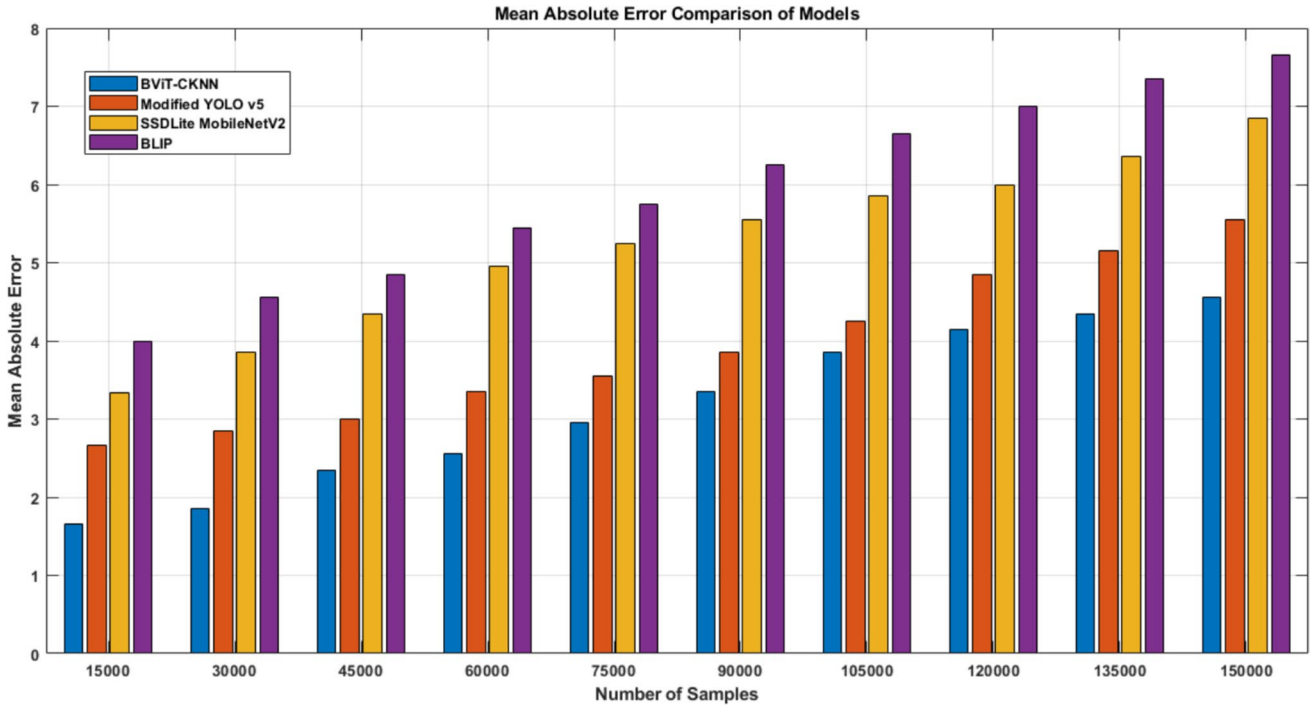

Fig. 12 Impact of Mean absolute error with the number of samples for proposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP

Table 6 Ablation study results of the proposed BViT-CKNN method compared to the baseline models for the COCO dataset

Methods/ Parameters name	Proposed BViT-CKNN	Existing modified YOLO v5 neural network [1]	Existing SSDLite MobileNetV2 [2]	BLIP
Precision	0.93	0.88	0.79	0.73
Recall	0.91	0.82	0.78	0.74
F1-score	0.92	0.85	0.78	0.73
Mean Square Error	3.16	3.9	5.23	5.95

compared to the existing methods [1, 2] using the COCO dataset.

7 Evaluation metrics for BLEU, METEOR, or CIDEr using proposed BViT-CKNN method

BLEU (Bi-Lingual Evaluation Understudy) is the evaluation metric to estimate the unigram or n-gram between the two images. It is a precision measure.The metric for Evaluation of Translation with Explicit Ordering (METEOR) is based on the harmonic mean of precision and recall.Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is the metric employed for the NLP domain. It related to image and video captioning evaluation based on the recall and precisionConsensus-Based Image Description Evaluation (CIDEr) is image and video description evaluation metric based on human consensus.The BLEU and METEOR are metrics mainlyconsidered for evaluating the quality of machine-generated text, such as in machine translation or image captioning. They measure the similarity between a candidate text and one or more reference texts. These metrics are not directly applicable to evaluating the accuracy of bounding box detection or classification in a visual pipeline. However, it is essential to recognize that in these cases, BLEU and METEOR are evaluating the linguistic output, and any correlation with bounding box or classification accuracy is an indirect consequence of the visual pipeline's contribution to the quality of the generated text. They are not direct metrics for evaluating the visual components themselves.

Figure 13 shows the comparison of BLEU, METEOR, or CIDEr for the proposed BViT-CKNN method and existing methods.

Figure 13 given above shows the BLEU, METEOR, or CIDEr results using four methods, namely proposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSDLite MobileNetV2 [2] and BLIP. As a result, the proposed BViT-CKNN provides better BLEU, METEOR, or CIDEr score values than compared to existing methods.

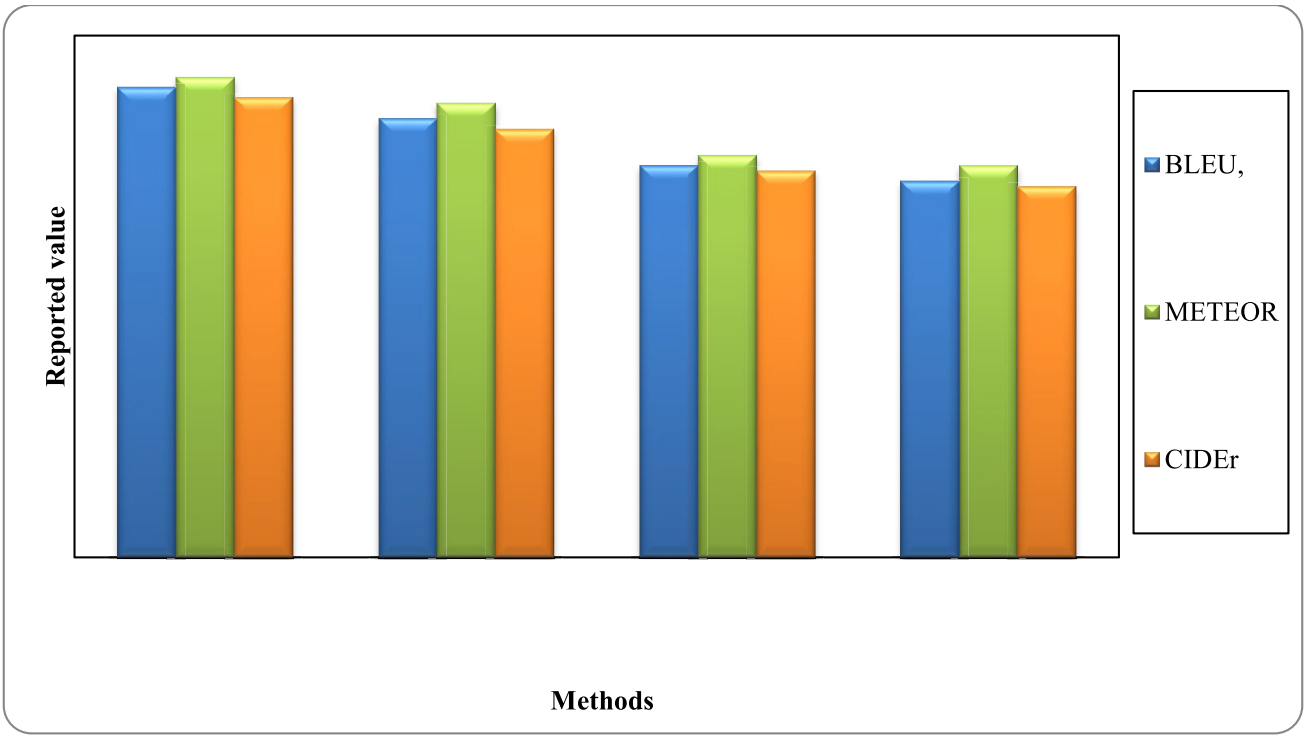

Fig. 13 Results of BLEU, METEOR, or CIDEr for proposed BViT-CKNN and existing methods, modified YOLO v5 neural network [1], SSD Lite MobileNetV2 [2] and BLIP

8 Real-world scenario of traffic crossings

In our work, the proposed BViT-CKNN is suitable for real-world traffic crossing scenarios. An obstacle detection system is utilized to improve safety. It is designed to recognize and signal the presence of crosswalks, helping users navigate to and cross them safely. It is used to identify and warn about potential hazards like bollards, poles, or parked vehicles that obstruct pedestrian traffic. Timely warnings are offered to prevent accidents by sudden obstacles in the environment. These systems significantly reduce the risk of collisions and injuries by determining the obstacles and providing alerts.The ability to detect and avoid obstacles can lessen the anxiety and fear associated with navigating public spaces.

9 Conclusion

Enhanced obstacle detection for visually impaired persons employing a Bilateral Vision Transformer and Convolution Kernel Neural Network (BViT-CKNN) is introduced.Preprocessing was designed withtwo processes for enhanced obstacle detection in the distinct collection of images sub-grouped into dissimilar classes. The preprocessing and global feature extraction section is analytical and straightforward, using a Bilateral Filter function and Vision Transformer-based Global Feature Extraction Generative Adversarial Network via Bounding Intersection Over Union-based Multi-Head Self-Attention aids in enhancing precision and recall. Obtained filtered images are fed into the Convolution Kernel Neural Network to carry out fine-grained local feature extraction for enhanced obstacle detection in dissimilar classes with improved accuracy. Finally, with the aid of the results learnt in the classification head activated via the sigmoid activation function, obstacle detection in distinct classes of objects was attained precisely and accurately. The outcome of BViT-CKNN is superior to that of traditional methods. The results of the proposed BViT-CKNN are to provide better performance with higher precision, recall, and F1 score with reduced MAE.

Acknowledgements The authors extend their appreciation to the King Salman center For Disability Research for funding this work through Research Group no KSRG-2024-318.

Author Contribution A.A. and M.S.A. conceptualized the research and designed the methodology. A.A. implemented the BViT-CKNN framework and conducted the experiments. A.A. and A.A.M. analyzed the results and contributed to performance evaluations. A.A. and A.A.A. prepared figures and tables. I.A. contributed to the literature review, manuscript structuring, and critical revisions. All authors contributed to writing the manuscript and reviewed and approved the final version.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Compting Interests The authors declare no competing interests.

References

1. Atitallah, A.B., Said, Y., Atitallah, M.A.B., Albekairi, M., Kaaniche, K., Boubaker, S.: An effective obstacle detection system using deep learning advantages to aid blind and visually impaired navigation. Ain Shams Engineering Journal **15**(2), 102387 (2024)
2. Islam, R.B., Akhter, S., Iqbal, F., Rahman, M.S.U., Khan, R.: Deep learning based object detection and surrounding environment description for visually impaired people. Heliyon (2023). https://doi.org/10.1016/j.heliyon.2023.e16924
3. Madake, J., Bhatlawande, S., Solanke, A., Shilaskar, S.: A qualitative and quantitative analysisof research in mobility technologiesfor visually impaired people. IEEE Access (2023). https://doi.org/10.1109/ACCESS.2023.3291074
4. Tarik, H., Hassan, S., Naqvi, R.A., Rubab, S., Tariq, U., Hamdi, M., Elmannai, H., Kim, Y.J., Cha, J.H.: Empowering and conquering infirmity of visually impaired using AI-technology equipped with object detection and real-time voice feedback system in healthcare application. CAAI Trans. Intel, Tech (2023)
5. Xia, K., Li, X., Liu, H., Zhou, M., Zhu, K.: IBGS: a wearable smart system to assist visuallychallenged. IEEE Access (2022). https://doi.org/10.1109/ACCESS.2022.3193097
6. Aung, M.M., Maneetham, D., Crisnapati, P.N., Thwe, Y.: Enhancing Object Recognition for Visually Impaired Individuals using Computer Vision. International Journal of Engineering Trends and Technology **72**(4), 297–305 (2024)
7. Ranganayaki, J., Iyonstan, P., Barath, S., Rajkumar, D., Sandeep, A.: Obstacle avoidance for blind people using machine learning, international journal of novel research and development, vol. 9, (2024)
8. Kharat, P., Kumar, T., Sirsikar, R., Sawant, R., Avhad, V.: Obstacle detection for visually impaired using computer vision. (2023)
9. Zhao, J., Liu, Z., Liu, Z., Xu, A., Zhao, J.: A lightweight blind obstacle detection network for mobile side. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **48**, 577–584 (2024)
10. Pydala, B., Kumar, T.P., Baseer, K.K.: Smart_Eye: a navigation and obstacle detection for visually impaired people through smart app. Journal of Applied Engineering and Technological Science (JAETS) **4**(2), 992–1011 (2023)
11. Maurya, A., Verma, P.: Optimized deep CNN-based obstacle detection for aiding visually impaired persons, revistainvestigationoperational, vol. 45, (2024)
12. Wang, W., Jing, B., Yu, X., Sun, Y., Yang, L., Wang, C.: Yolo-od: Obstacle detection for visually impaired navigation assistance. Sensors. **24**(23), 7621 (2024Nov 28)
13. Xu, Y., Khan, T.M., Song, Y., Meijering, E.: Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. Artif. Intell. Rev. (2025). https://doi.org/10.1007/s10462-024-11033-5
14. Schiatti, L., Gori, M., Schrimpf, M., Cappagli, G., Morelli, F., Signorini, S., Katz, B., Barbu, A.: Modeling visual impairments with artificial neural networks: a review. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1987-1999) (2023)
15. Wang, J., Wang, S., Zhang, Y.: Artificial intelligence for visually impaired. Displays **77**, 102391 (2023)
16. Masud, U., Saeed, T., Malaikah, H.M., Islam, F.U., Abbas, G.: Smart Assistive System for Visually Impaired People Obstruction Avoidance through Object Detection and Classification. IEEE access **10**, 13428–13441 (2022)
17. Feng, S., Wang, Y., Gong, J., Li, X., Li, S.: A fine-grained recognition technique for identifying Chinese food images. Heliyon (2023). https://doi.org/10.1016/j.heliyon.2023.e21565
18. Lu, J., Zhang, W., Zhao, Y., Sun, C.: Image local structure information learning for fine-grained visual classification. Sci. Rep. **12**(1), 19205 (2022)
19. Asiedu Asante, B.K., Imamura, H.: Towards Robust Obstacle Avoidance for the Visually Impaired Person Using Stereo Cameras. Technologies **11**(6), 168 (2023)
20. Hoang, V.N., Nguyen, T.H., Le, T.L., Tran, T.H., Vuong, T.P., Vuillerme, N.: Obstacle detection and warning system for visually impaired people based on electrode matrix and mobile Kinect. Vietnam Journal of Computer Science **4**(2), 71–83 (2017)
21. Li, D., Zhang, J.: Intelligent road segmentation and obstacle detection for autonomous railway vehicle. Advances in Mechanical Engineering **16**(1), 1–11 (2024)
22. Feng, H., Liu, W., Xu, H., He, J.: A lightweight dual-branch semantic segmentation network for enhanced obstacle detection in ship navigation. Eng. Appl. Artif. Intell. (2024). https://doi.org/10.1016/j.engappai.2024.108982
23. Bamdad, M., Scaramuzza, D., Darvishy, A.: SLAM for Visually Impaired People: A Survey. IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3454571
24. Valipoor, M.M., De Antonio, A.: Recent trends in computer vision-driven scene understanding for VI/blind users: a systematic mapping. Univ. Access Inf. Soc. **22**(3), 983–1005 (2023)
25. Ikram, S., Sarwar, I., Ikram, A., Abdullah-AI-Wahud, M.: A Transformer-Based Multimodal Object Detection System for Real-World Applications, IEEE Access, vol. 13, (2025)
26. Zhang, W., Chai, Q., Zhang, Q., Wu, C.: Obstacle-transformer: A trajectory prediction network based on surrounding trajectories. IET Cyber-Systems and Robotics (2023). https://doi.org/10.1049/csy2.12066
27. Li, X., Huang, M., Xu, Y., Cao, Y., Lu, Y., Wang, P., Xiang, X.: AviPer: assisting visually impaired people to perceive the world with visual-tactile multimodal attention network. CCF Transactions on Pervasive Computing and Interaction **4**(3), 219–239 (2022)
28. Khan, W., Hussain, A., Khan, B.M., Crockett, K.: Outdoor mobility aid for people with visual impairment: Obstacle detection and responsive framework for the scene perception during the outdoor mobility of people with visual impairment. Expert Syst. Appl. **15**(228), 120464 (2023Oct)
29. Kuriakose, B., Shrestha, R., Sandnes, F.E.: DeepNAVI: A deep learning based smartphone navigation assistant for people with visual impairments. Expert Syst. Appl. **212**(1), 118720 (2023)

30. Leong, X., Ramasamy, R.K.: Obstacle detection and distance estimation for visually impaired people. IEEE Access (2023). https://doi.org/10.1109/ACCESS.2023.3338154
31. Said, Y., Atri, M., Albahar, M.A., Ben Atitallah, A., Alsariera, Y.A.: Obstacle Detection System for Navigation Assistance of Visually Impaired People Based on Deep Learning Techniques. Sensors **23**(11), 5262 (2023)
32. Bayat, N., Kim, J.H., Choudhury, R., Kadhim, I.F., Al-Mashhadani, Z., Virgen, A.D., M., Latorre, R., De La Paz, R. and Park, J.H.: Vision Transformer Customized for Environment Detection and Collision Prediction to Assist the Visually Impaired. Journal of Imaging **9**(8), 161 (2023)