A Transformer-Based Multimodal Object Detection System for Real-World Applications

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 January 2025, accepted 4 February 2025, date of publication 6 February 2025, date of current version 18 February 2025.

*Digital Object Identifier 10.1109/ACCESS.2025.3539569*

RESEARCH ARTICLE

# A Transformer-Based Multimodal Object Detection System for Real-World Applications

**SUNNIA IKRAM[1], IMRAN SARWAR BAJWA[1], AMNA IKRAM[2], M. ABDULLAH-AL-WADUD[3], (Member, IEEE), AND HALEEMA PK[4,5]**

[1]Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[2]Department of Computer Science, The Government Sadiq College Women University of Bahawalpur, Bahawalpur 63100, Pakistan
[3]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4]School of Computing and Engineering, University of West London, RAK Branch Campus, Ras Al-Khaimah, United Arab Emirates
[5]Rochester Institute of Technology, Rochester, NY 14623 USA

Corresponding author: Sunnia Ikram (sunnia.ikram@iub.edu.pk)

This work was supported by the King Saud University, Riyadh, Saudi Arabia, under Grant RSPD2025R951.

**ABSTRACT** Obstacle detection is a critical task for visually impaired individuals to ensure safe navigation and hazard avoidance. This study presents FusionSight, an innovative multimodal fusion model that integrates radar and image data to address challenges in real-time object classification for dynamic environments. The system leverages an Arduino Uni microcontroller for data acquisition and transmission, enabling seamless communication between radar and image datasets and the cloud environment. For image data, the Vision Transformer (ViT) was employed to extract high-level features, capturing fine details and long-range dependencies essential for accurate object recognition. Concurrently. Radar data was processed using a Convolutional Neural Network (CNN) to extract spatial and temporal features such as distance, speed and velocity critically for understanding object dynamics. To unify these diverse modalities, a Feature Fusion Multimodal Transformer (FFMA) was utilized, facilitating the integration of complementary features into a comprehensive representation. This fusion mechanism enables the model to effectively handle challenges such as occlusion, overlapping objects and varying lighting conditions. The unified features were classified into four categories ''close, far, moving and fast-moving'' using a Feed-Forward Neural Network (FFN). The classification results were then converted into actionable audible feedback, providing real-time navigation assistance to visually impaired users. The FusionSight model set a benchmark in multimodal data fusion, achieving an impressive classification accuracy of 99% when using static dataset and 98% accuracy with real-time dataset. This study demonstrates the practical implementation navigation for visually impaired individuals and other use cases involving dynamic and complex environments.

**INDEX TERMS** Obstacle detection, visual impairment, mini camera, radar sensor, CNN, FFMT, ViT, FFN.

## I. INTRODUCTION

Obstacle detection is a common function required in many fields such as automotive, robotics and surveillance, amongst others. In recent years, the field has evolved significantly, benefiting industries such as autonomous systems, robotics and assistive technologies. Among these applications obstacle detection and classification continue to be important challenges, especially for systems aimed at improving safety and navigation.

Approximately, 39 million people are completely visually impaired, and it is expected that the number of visually impaired individuals will double by 2024 [1]. These individuals face numerous difficulties in socializing and completing daily tasks, often requiring assistive devices for safer navigation. Nunes, in [2] by Nunes, D., et al worked on the challenges visually impaired individuals face while walking and addressed the problem in a real-time environment by recognizing, detecting and identifying objects. Users need devices that help them navigate safely to avoid collisions. Such devices trained with machine learning and artificial intelligence models can offer significant independence [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Dominik Strzalka.

Advanced technologies have been employed in obstacle detection systems, including sensors and cameras. These systems, as demonstrated in [4], utilize real-time image processing to detect objects and integrate text-to-speech functionalities for audio feedback [5], [6]. Ultrasonic sensors, which are popular for their low cost and short-range detection capabilities, use sound waves to measure distance, offering a simple and cost-effective alternative [7].

However, some challenges remain in the previous work, for example in the detection of objects and providing real-time navigation assistance. Radar sensors are useful in these situations and some of the benefits include long range detection and speed estimation. However, radar sensors are often regarded as costly and sensitive to weather conditions, while economical substitutes include ultrasonic sensors and monocular cameras in low-cost applications. This flexibility makes such solutions versatile to a broad client base.

Although there are methods for both object detection and speed estimation, they are based on single-modality data, which are more sensitive to realistic conditions such as low lighting or changing road conditions. These limitations are overcome in the FusionSight model that we proposed here, which uses a multimodal transformer to combine radar and camera data. This ensure improved accuracies and reliability of the algorithm in identifying, categorizing and positioning the obstacles for navigation aid [6], [8]. The fusionSight approach employs several efficiency enhancements including preprocessing that involves normalization, augmentation and noise correction that guarantee high quality data with low computational complexity. Moreover, the real-time sensor data fusion in lightweight architectures enables the system to be implemented on mobile devices, thus being both feasible and efficient.

To overcome these challenges, we suggest the use of multimodal data fusion for enhancing obstacle detection and classification. This method integrates image data with radar data since the two data types have their strengths. Image gives detailed picture information while radar gives detailed distance and speed information. Combining these data types enhances the system's environmental comprehension, which is essential for obstacle detection classification. The core of this approach is based on the multimodal transformer model that combines image and radar data. The integration of ViT for detailed image processing and CNN for radar data processing guarantees that the feature extraction process is efficient. These methods help to decrease the computational cost inherent to multimodal fusion, which allows for real-time processing. The fused data is then passed through a FFN which categorizes objects into close, far, moving and fast-moving.

This approach is expected to improve obstacle detection and classification accuracy by utilizing the unique strengths of radar and camera modalities. Furthermore, the adaptability of the model allows its deployment with cost-effective alternatives, making it accessible for visually impaired individuals and other use cases.

The primary aspect of this study is that highlight the novel aspects of the work are as follows:

- The application of transformers for multimodal fusion in real-time obstacle detection is a novel contribution that effectively handles diverse sensor data types.
- This work uses ViT for extracting images features in the context of assistive technologies for the visually impaired. This integration enhances the ability of the model to capture detailed visual information, which is crucial for accurate obstacle detection.
- The use of CNN to extract features from radar data effectively captures spatial and temporal patterns which have not been extensively explored in obstacle detection systems.
- The proposed system incorporates optimization techniques to improve computational efficiency while maintaining high accuracy in classifying obstacles. These enhancements ensure the feasibility of real-time deployment in dynamic, resource-constrained environments.
- The model achieves unprecedented accuracy in classifying obstacles into different categories. This high accuracy set has not been extensively explored a new standard for multimodal classification systems in dynamic, real-time environments. The ability to reliably fuse and classify diverse sensor data in real-time is a significant advancement.

The remainder of this paper is arranged as follows:

Section II introduces the theoretical foundations. Section III details the proposed methodology. Section IV presents the experimental verification. Section V summarizes the conclusion.

## II. RELATED WORK

Visually impaired people face several challenges while walking on the roads or indoor environment. Most of the time, they feel dependent on others, this is a major problem for them walking in an unfamiliar environment [9]. This research gives an image-based solution to the problem by using RGB-D cameras, which supply scene images with a balance between accuracy and sensor cost. This is why detection accuracy depends heavily on the sensors utilized and the processing of the information has a significant impact on efficiency [10], [11]. This obstacle detection model employs a respeak synthesizer on a smartphone to generate the audio output for the user. The model demonstrated a high-speed assistive device with high accuracy [12].

Techniques such as those discussed in [8] focus on straightforward calculations of distance and speed using single-modality data. While effective for basic tasks, these approaches lack the ability to handle complex, real-world scenarios where multimodal data integration is required for improved spatial localization and classification. The FusionSight model addresses these challenges by employing advanced data fusion mechanisms, which enable better object detection and classification in dynamic settings.

**TABLE 1.** A Literature Review on various sensors and technologies contributions to address different challenges in Obstacle Detection Systems.

| References | Hardware | Fusion and Classification Technologies | Feature Extracted | Achievements |
|---|---|---|---|---|
| 2023, [13] | Infrared cameras, RGB cameras, and Light detection and ranging (LiDAR) sensors | Gaussian (DoG), YOLO-v5, multisensory fusion | Semantic information, 3-D coordinates of objects, working environment at night. | Accurately extracting pedestrian position information, realizing timely pedestrian alarms |
| 2023, [14] | LiDAR, a camera, and a GNSS/INS, | Focal Voxel R-CNN, sparse convolution | Extract effective features | Detection for small objects, high precision increased from 89.04% to 92.89% |
| 2024, [15] | (LiDAR) system | global planar features, Iterative Closest Point, filtering algorithm | Imaging range, angular resolution, extract geometric edge features | High-precision imaging within a 50 m monitoring area, 70 % detection rate for obstacles |
| 2024, [16] | Advanced Ultrasonic Motor | Pulse Code Modulation | User's language preferences, warning, High autonomy, Safety | Detecting elevated obstacles, based on the distance. |
| 2024, [17] | EEG signals, wheelchair | Two convolutional neural networks (CNNs), brain-machine interface | Imagery classification | Accuracy of 83% in classifying EEG signals, collision avoidance |
| 2024, [18] | AI-driven outdoor obstacle detection | YOLOv5, Google Text-to-Speech | Visual impairments in outdoor environments, object recognition | High accuracy in obstacle detection |
| 2024, [19] | 3D printed wearable smart glass, Ultrasonic Sensors | Regression model | Guide individuals in an indoor space | Detecting obstacles with an F1-score of 84.7% (at 100% precision) |
| 2023, [20] | Ultrasonic sensors | voice commands | Minimum width's distance | Higher accessibility, comfort, and simplicity of navigation |
| 2023, [21] | Ultrasonic Sensors Kinect Sensor, Local Routing | Convolutional Neural Network , Google Voice | Object identification, spoken words, Feature Maps | Surroundings for obstacles, spoken words |

The techniques used in previous works to address Obstacle detection challenges in various environments are summarized in Table 1. For night-time environments, infrared cameras have been utilized for detection tasks despite their lower resolution compared to EGB cameras.

To mitigate this, a difference of Gaussian (DoG)-based image enhancement method is applied to low-resolution infrared images; subsequently, a YOLOv5-based images; recognition method is utilized for enhanced image detection. Finally, a multi-sensor fusion technique identifies semantic information and 3D coordinates of objects. Experimental results demonstrate the method's capability to accurately detect pedestrian positions and issue timely alarms, maintaining performance in nighttime excavator operations [13].

This study applies deep learning to LiDAR-based 3D obstacle detection for landscapes, using a data acquisition platform with LiDAR, a camera, and GNSS/INS on agricultural machinery as shown in Table 1. An effective 3D obstacle detection method, Focal Voxel R-CNN, is trained using focal sparse convolution to enhance feature extraction from sparse point clouds. This approach improves detection performance for small objects, with pedestrian detection AP increasing from 89.04% to 92.89%, and an overall mAP of 91.43%, which is 3.36% higher than the base model [14].

In [15] by Zhu, G., et al, a LiDAR system was designed for railway environments, capable of autonomously adjusting imaging range and angular resolution to achieve high-precision imaging within a 50-meter monitoring area. Using registration and filtering algorithms, a railway surface

obstacle extraction algorithm achieves a 100% detection rate for obstacles of at least 15cm height and a 70% detection rate for 10cm obstacles. This system features sub-centimeter registration accuracy and precise geometric extraction enhancing railway safety. In [16] by Păpară, R., et al, focus was on developing a user-friendly and cost effective obstacle detection system that provides multi-language support and eliminates the need for IT expertise. The system enhances safety and accessibility by delivering obstacle detection alerts based on proximity.

Additionally, an object detection system integrates vision data with a brain-machine interface to classify motor imagery using a modified 10-20 electrode setup and convolutional neural networks, achieving 83% accuracy in EEG signal classification and enabling real-time robot control for collision avoidance. In [18] by Bougheloum, L, an AI-driven system supported visually impaired individuals in indoor navigation using YOLOv5 for object recognition and Google Text-to-Speech for audio feedback. Trained on a customized dataset and the MS COCO dataset, the model provides high accuracy in obstacle detection and converts detection results into real-time audio feedback via earphones, significantly enhancing outdoor mobility for visually impaired users. In [19] by Kusneniwar, H.G, Sonic Glass a wearable indoor navigation system, incorporated multiple sensors and microcontrollers in a 3D-printed smart glass. User studies indicate Sonic Glass effectively guides users indoors with an 84.7% F1-score in obstacle detection and a 1.35-meter localization error, offering valuable assistance to visually impaired individuals.

In [20] by J, D., et al, proposed a project to accurately measure the minimum width distance using ultrasonic sensors. The sensors measure distances, and the device responds to verbal commands. In [21] by Shahani, S. and N. Gupta, conducted a study on navigation systems for visually impaired individuals who need warnings when they encounter obstacles in their path. TensorFlow is used for obstacle detection when implementing deep learning algorithms. This research focuses on the acquisition of real-time data to check the performance of the model. In this context, the proposed research will address the following research question:

This project aims to determine precise distance from various obstacles, making it applicable in areas such as robotics, car sensors for obstruction avoidance, and distance measurements in construction sites. The development of the device incorporates multiple applications from university courses, including Microprocessor, Basic Engineering Multimedia, and Electronics. The result of this study shows improved accuracy, the device was more comfortable for visually impaired individuals and allowed for easier compared to the traditional white cane, with the audible output produced for the user.

The following are the research questions:

RQ1: What are the benefits of integrating image and radar data in obstacle classification compared to using only one of them?

RQ2: How do ViT perform in terms of feature extraction and image processing, and how efficient are they compared to convolutional image processing techniques?

RQ3: How does multimodal transformer architecture help in the fusion of image and radar features and what are the advantages of this approach over simple fusion?

A refined cross-check method is employed to ensure high-quality results without compromising real-time performance. In summary, many scholars have conducted extension work, and significant progress has been made on radar and image data fusion. However, several challenges remain unsolved, including effective joint calibration and mitigating the impact of image distortion.

## III. MATERIAL AND METHODS

The FusionSight model was developed to overcome the challenges of dynamic, real-time obstacle detection for visually impaired users. Unlike simpler methods, which rely on single-sensor data, FusionSight integrates radar and image features using FFMT. This enables the model to handle overlapping objects, occlusions and varying lighting conditions, providing a richer and more reliable representation of the environment for enhanced classification and localization.

The modular design of the FusionSight model allows for hardware flexibility. For cost-sensitive applications, LiDAR can be replaced with low-cost ultrasonic sensors or monocular cameras, albeit with potential trade-offs in detection accuracy and range.

The study employs two distinct datasets, one of which contains images while the other contains radar sensor data. The image dataset includes varied objects from outdoor environments, whereas the radar dataset contains sensor data related to the image scenarios, providing additional information such as distance, speed, motion, place, longitude, latitude, date, time and object location. Both datasets were sourced from Kaggle to train the model and were tested on real-time datasets. Each of the datasets was in zip files.

The architecture of the FusionSight model, depicted in Figure 1, showcases how image and radar datasets are transmitted to the cloud for processing using an Arduino UNO microcontroller. This microcontroller serves as an interface for data acquisition and real-time transmission, ensuring seamless communication between the data sources and the cloud environment for further analysis. The implementation of the FusionSight model was developed using Python 3.10 framework was utilized, providing robust support for model design, training and optimization. To handle feature extraction and classification, the Transformer Library was leveraged, specifically employing the ViT model for processing image data. The ViT model efficiently extracts high-level features from images, enabling the system to identify objects with increased accuracy.

The integration of ViT for visual data and CNN for radar data represents a novel approach to multimodal fusion. ViT excels at capturing long-range dependencies in image data, making it highly effective for identifying subtle details in

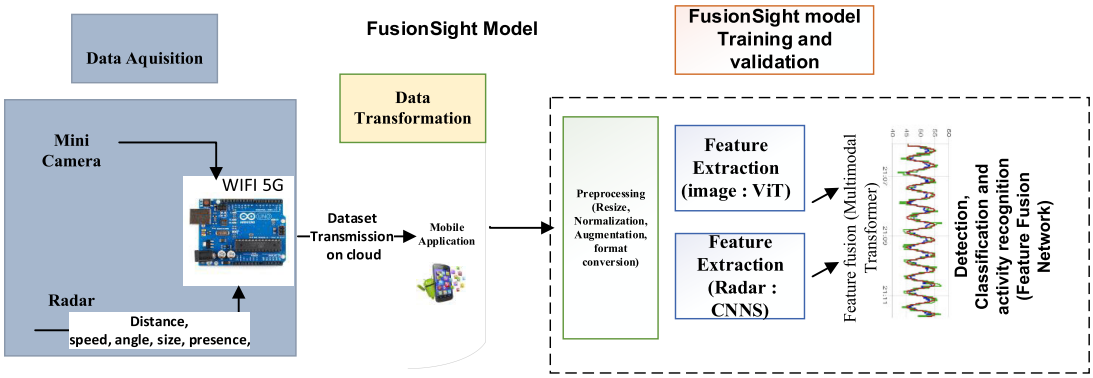


**FIGURE 1.** Architecture of FusionSight model.

complex environments. Conversely, CNN is adept at extracting spatial and temporal patterns from radar data, such as distance and velocity. This combination ensures that the model benefits from the complementary strengths of both modalities, providing a robust understanding of the environment. Unlike previous approaches, which focused on a single modality or lacked effective integration techniques, the FusionSight model employs a multimodal transformer architecture to unify these features into a comprehensive representation. This innovative fusion mechanism greatly improves the model's ability to classify obstacles in dynamic and real-world scenarios.

For radar data, CNN was used to extract features from the radar data since it was designed to handle data from the radar sensor by identifying patterns that are associated with distance, speed and other radar metrics. The features extracted from both image and radar modalities were then fused using FFMT to allow the integration of the complementary data sources into a single feature space. This model was specifically designed to process data from the radar sensor, and extract patterns concerning distance, speed and other radar parameters. The fused representation was classified into four distinct categories: close, far, moving and fast-moving using the FFN. The FFN then sums up the processed features to come up with the final output of the classification. The results of the classification were then translated into an audible form which could be used to give feedback to the user in real-time, especially to visually impaired to navigate and understand the environment around them.

### A. DATA ACQUISITION AND PREPARATION

Prior to the model training, the datasets required preprocessing. The first step involved extracting the image and radar datasets from their respective files. Since these datasets were initially unlabeled, appropriate labels were generated to categorize the data into one of four predefined classes: "close, far, moving and fast-moving" [11]. These labels were critical for subsequent classification tasks and were assigned based on radar metrics such as speed, distance and object trajectory.

To improve data quality while maintaining computational efficiency, a combination of normalization noise reduction and augmentation techniques was applied.

- Radar data values were scaled to ensure uniformity and eliminating irrelevant features and preserving critical metrics such as speed, distance and angle. This was essential to ensure accurate feature extraction from the radar data.
- Noise in radar signals was minimized using low-pass filter, ensuring that only relevant features were extracted. This step was particularly important to maintain the fidelity of radar metrics such as speed and distance.
- Data Augmentation techniques such as rotation, flipping and zooming were applied to the image dataset to increase its diversity. This step not only improved the model's generalization but also ensured its robustness in various real-world scenarios.

Initially the model was trained on the datasets available on Kaggle, and later it was experimented in real-time environment and achieved high accuracy in both datasets. The experiments have been done by taking dataset of outdoor environment where user face traffic and it was trained for the same environment.

#### 1) REAL-TIME DATA COLLECTION

To collect real-time data, a mini camera and radar module were attached to the cap of the user. These sensors captured live data, which was transmitted to a mobile application via a microcontroller and Wi-Fi 5G.

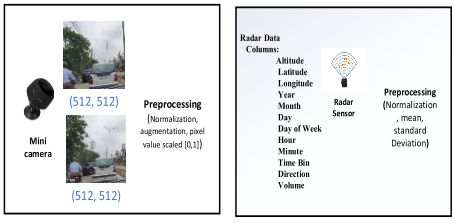


**FIGURE 2.** Data acquisition and preprocessing steps used in FusionSight model.

The mobile application served as an intermediary, transmitting the collected data to the cloud for further preprocessing and analysis [11], [22]. Data preprocessing is used in the preparation of input data for the multimodal fusion model. This process comprises of normalization, augmentation on

images, and formal conversion on radar datasets. It was crucial to normalize the pixel value before passing the image data into ViT. The images are represented as tensors with pixel values ranging from 0 to 255. The pixel values are normalized to the range [0, 1] using this equation. Figure 2 illustrates the data acquisition and preprocessing steps of the FusionSight model. The figure shows how data from radar and images are collected, transmitted and prepared for input into the multimodal fusion model. This streamlined pipeline plays a crucial role in the model's ability to accurately classify objects in diverse environments.

$$I_{norm}(x, y, z) = \frac{I(x, y, z)}{255}, \tag{1}$$

where,

- I(x, y, z) represents the pixel value at position (x,y,z) in the c color channel (red, green, blue)
- $I_{norm}$ is the normalized pixel value.

The radar data contains continuous numerical values speed, distance; direction is normalized to ensure that all features contribute equally to the model's learning process. Equation used for normalization of the radar dataset is:

$$R_{norm}(i) = \frac{R(i) - R_{min}}{R_{max} - R_{min}}, \tag{2}$$

where,

- R(i) represents the i-th feature in the radar data.
- $R_{min}$ and $R_{max}$ are the minimum and maximum values of the feature R(i).
- $R_{norm}$ is the normalized feature value.

The overall architecture of the model is integrated with 4 major phases as illustrated in Figure 2. The main part of this proposed model is the model training and validation section. (1) Data acquisition and preparation, (2) Feature extraction by using two different feature extraction models ViT and CNN, (3) FFMT is used for combining both data, (4) FFN for classification and recognition.

### B. FEATURE EXTRACTION OF BOTH DATASETS

The FusionSight model employs a layered approach for feature extraction, tailored to the distinct characteristics of image and radar data, ensuring an optimal combination of advanced deep learning techniques [23].

For image data, ViT is utilized due to its superior ability to capture global attention patterns. ViT segments the input image into patches, encodes them and applies a self-attention mechanism to find relevant patterns throughout the image as shown in Figure 3.

This approach provides a strong foundation for feature extraction and is especially useful in situations where high-resolution spatial context is needed, for instance, in recognizing intricate visual objects in the environment.

Radar signals are digital data that are inherently spatial and temporal in nature. CNNs, with their convolutional filters are effective at extracting these features, which in essence, model strength the strength of signals, position, velocity and

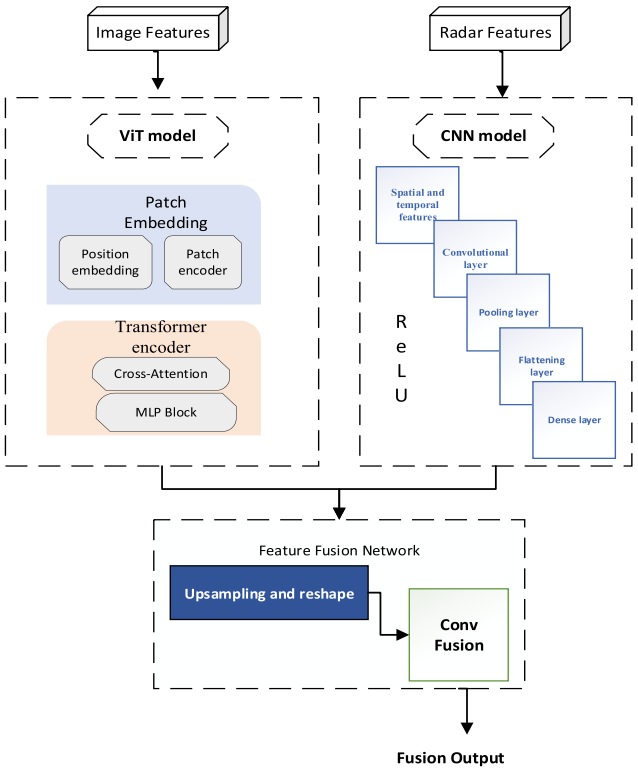


**FIGURE 3.** Layered approaches used by FusionSight model for feature extraction.

movement. While ViT is developed for unstructured image pixels, CNN offers a computationally efficient and effective way to process radar signals.

The radar data is preprocessed into structured representations so that CNN can extract the temporal-spatial features. These features are then fused using a multimodal Transformer, ensuring both modalities provide unique and complementary information to the final classification task. This layered methodology is vital for accurately classifying objects into four categories: close, far, moving and fast-moving. The integration of radar and image data through this design allows the model to enhance its performance for real-time obstacle detection and object classification, specifically tailored to the needs of visually impaired users.

The formally description of process is:

1. The image is divided into smaller patches, each of size p*p and image I of size H*W, this results in

$$N = \frac{H \times W}{P_2}, \tag{3}$$

- N is the total number of patches the image is divided into.
- H is the height of the input image
- W is the width of the input image.
- P is the patch size (typically square, with dimension P*P)

2. Each pitch is flattened and linearly projected into an embedding space:

$$Z_i = W \times Flatten(I_{i,j}) + b, \tag{4}$$

where,

- W is the weight matrix of dense layer.
- Flatten (Ii) is the input Ii after being flattened into a one-dimensional vector.
- b is the bias vector which allows the model to shift the activation function output.
- $Z_i$ is the output of the dense layer for input $I_i$.

3. The embedded patches are passed through transformer layers to capture global dependencies denoted as:
- Z‘ = Transformer (Z).
- Z = [$z_1$, $z_2$, $z_3$, ….] is the sequence of patch embeddings. Z‘ is the sequence of processed embeddings after the transformer layers.

4. A CNN-based approach is used, applying convolutional layers to extract relevant features from the radar sensor inputs [24]. The process used can be describe mathematically as:

$$\boldsymbol{F(i,j)} = \sum_{\boldsymbol{m}=1}^{\boldsymbol{M}} \sum_{\boldsymbol{n}=1}^{\boldsymbol{N}} \boldsymbol{W_{m,n}} \times \boldsymbol{R(i+m-1, j+n-1)} + \boldsymbol{b} \quad (5)$$

where,

- $F_{(i,j)}$ is the value at (i,j)th position in the output feature map
- R(i+m-1, j+n-1) refers to values from the input matrix R, shifted by indices m and n. This represents a small ‘‘window’’ or kernel sliding over the input matrix
- M is the size of the kernel or filter in the vertical direction (height).
- N is the size of the kernel or filter in the horizontal direction (width).
- m and n indices used to iterate over the kernel dimensions.
- b is the bias term added to the result, often used in convolutional layers to adjust the output.

The CNN layers capture spatial hierarchies in the radar data, producing feature maps that summarize the relevant information for classification.

### C. FEATURE FUSION (MULTIMODAL TRANSFORMER)

The extracted features from both image and radar data are then combined through a feature fusion network. This fusion network plays a critical role in ensuring that the joint representation of the two data modalities is effectively learned. This approach is intended to work with the relationship between different modalities (radar and image) and integrate them into a single representation that improves the classification performance. Both radar and image features are enriched with positional encodings to preserve the sequence order in which the data is processed.

$$Fr = Fr + PosEnc, \quad (6)$$

where,

- Fr is the feature representation of the input patch.
- PosEnc denotes the positional encoding matrix added to the features to encode positional information.

The addition operation ensures that the model differentiates between elements based on their position.

$$Attention\,(Q, K, V) = SoftMax\left(\frac{Qk^T}{\sqrt{d}_K}\right), \quad (7)$$

where,

- Q, K and V matrices derived from the input sequence or data:
- Q represents the vector that asks, ‘‘what am I looking for?’’
- K represents the vector that encodes ‘‘what information is available?’’
- V represents the vector that contains the ‘‘actual content or information’’. They are obtained by multiplying the input sequence with learned weight matrices.
- $QK^T$computes the dot-product similarity between radar and image features.
- $\sqrt{d_k}$is a scaling factor to prevent extremely large values in the dot-product.
- The SoftMax function normalizes these scores, highlighting the most relevant image features corresponding to each radar feature.

*SoftMax* converts scores into attention weights, emphasizing the most relevant features. It ensures that the attention mechanism focuses on the most important interactions between the image and radar features. Normalization via scaling and SoftMax ensures efficient learning and improved model performance. The SoftMax function normalizes these scores, highlighting the most relevant image features corresponding to each radar feature. The method used for Feature Fusion; the attention scores are used to create a fused features representation:

$$F_{fused} = Attention\,(Q, K, V), \quad (8)$$

where,

- $F_{fused}$ is the fused feature representation obtained after applying the attention mechanism.
- Q represents the current data’s feature vector, guiding the attention to focus on relevant part of the input.
- K encodes all input features, contextual information, determining their relevance to the query.
- V contains the actual data values to be weighted and summed based on the attention scores.

The attention mechanism computes a compatibility score between Q and K, normalizes it and uses it to weight V. It enhances the integration of multimodal data by aligning complementary information. Ensure that both image and radar data contribute meaningfully to the classification task, improving the model’s accuracy and robustness.

### D. FEATURE FUSION NETWORK FOR CLASSIFICATION

After the fusion of radar and image features, the final step is to classify the fused features into predefined categories such as ‘‘close, far, moving. Fast-moving’’. The classification process works after the fusion of radar and image features [25].

The extracted features from the FNN are then fed through one or more fully connected layers. These layers perform a linear transformation on the input features and then apply a non-linear activation function (ReLU) to the input features.

The mathematical function of the fully connected layer used in this model is:

$$H = ReLU(W \times F_{fused} + b), \tag{9}$$

where,

- H is activated feature representation after applying the ReLU function. H serves as the input for subsequent layers in the model.
- ReLU activation function is applied element-wise to the output of the linear transformation.
- W is the weight matrix,
- $F_{fused}$ represents the fused features obtained by combining information from multiple modalities.
- b is the bias vector added to the weighted features to ensure flexibility in the representation.

The output from the fully connected layer is then passed to the final classification layer, often a *SoftMax* layer. The *SoftMax* function transforms the output logits into probabilities of each class.

The output from the fully connected layer is then passed to the final classification layer, often a *SoftMax* layer. The *SoftMax* function transforms the output logits into probabilities of each class. The class with the higher probability is chosen as the final output of the model. The *SoftMax* layer gives out the probability distribution of all the classes and the class with the highest probability is the prediction class. Cross-entropy loss function is used to measure the difference between the predicted class probability and the actual class label.

$$\boldsymbol{Loss} = -\sum_{k=1}^{K} \boldsymbol{yklog}\left(\boldsymbol{p}\left(\boldsymbol{y} = \frac{\boldsymbol{k}}{\boldsymbol{h}}\right)\right) \tag{10}$$

where,

- Loss is the total loss value that needs to be minimized during model training.
- $\Sigma_{k=1}^{k}$ indicates summation over all K classes in the classification task.
- $y_k$ is the actual label for class k.
- $(P(y = k|h))$ is the predicted probability that the input belongs to class k, given the features h.
- $\mathrm{Log}(P(y = k|h))$ is the natural algorithm of the predicted probability for class k.

This equation defines the cross-entropy loss function used for classification tasks, measuring the difference between predicted probabilities and true labels. Loss is the overall loss value, which the model seeks to minimize during training. K is the total number of classes in the classification task. The yk is the ground truth label for class k, where $y_k = 1$, if the true class is k, and $y_k = 0$ otherwise. The $p(y = k|h)$ is predicted probability of class k, given the input features h, typically output by a *SoftMax* layer. log $(p(y = k|h))$ is the logarithm of the predicted probability, penalizing incorrect predictions more heavily. The parameters of the model are optimized during training to minimize the loss function. Gradient descent-based optimization technique Adam optimizer is used for optimization. The fused features are subsequently passed through a fully connected layer, where the final classification into one of the four categories is performed using ViT for image classification. Compared to the prior research that directly concatenates features from different modalities, the fusion module in our model does not favor any modality. However, it dynamically extracts relevant features from both radar and image data depending on the classification results that are required. In particular, the proposed method uses cross-attention where radar features are considered as the query and image features are considered as the key and value. This makes it possible for the model to selectively augment radar data with image features that are relevant for the fusion process, thus making the fusion process more effective and ensures a positive correlation between the features.

### E. TRAINING PROCEDURE

The model training procedure involves several steps. The loss function employed is Cross-Entropy Loss *(nn.CrossEntropyLoss())* which is suitable for multi-class classification tasks. The Adam optimizer, with a learning rate of 1e-4, is used to optimize the model parameters. The model was trained over 20 epochs, with loss and accuracy computed at each epoch to monitor the training progress. The training loop iteratively adjusted the model parameters to minimize the loss function while increasing classification accuracy.

## IV. EXPERIMENTAL EVALUATION

The experimental analysis of FusionSight model evaluates its effectiveness in obstacle detection and classification for visually impaired individuals. The experiments were designed to test the model's adaptability and performance across a range of scenarios, including controlled datasets and real-life conditions. Results indicate that while LIDAR technology provides the highest accuracy, alternative low-cost configurations (e.g. ultrasonic sensors and monocular cameras) can still deliver reliable performance for short-range applications, making the model feasible for budget-constrained users. This section describes the experiments conducted, the datasets used for the experiments, the evaluation metrics used, and the results achieved. Initially training and validation were conducted using a multimodal dataset comparing radar and image data from Kaggle. This dataset allowed controlled evaluation of the FusionSight model under well-defined conditions.

To validate the model's performance in practical applications, real-time datasets were collected using a mini camera and radar module attached to the user's cap. These sensors captured images and radar signals in outdoor environments, such as busy streets and obstacle-rich pathways. The real-time dataset included diverse scenarios like varying lighting conditions, weather changes, and dynamic obstacles, ensuring comprehensive evaluation of the model's robustness.

**TABLE 2.** Parameters used by ViT, CNN, FFT classification layer, training parameters in fusionsight model.

| Component | Parameter | Description | Value/Range |
|---|---|---|---|
| Vision Transformer (ViT) | Image-size | Size of the input images | 224*224 pixels |
| | Patch-size | Size of each patch | 16*16 |
| | Embed-dim | Embedding dimensions of patches | 768 |
| | Num-heads | Number of attention heads | 12 |
| | Depth | Number of transformer layers | 12 |
| | mlp-dim | Dimension of the mlp in transformer layers | 3072 |
| | Dropout-rate | Dropout rate in transformer layers | 0.1 |
| | Num-classes | Number of output classes | 4(close, far, moving, fastmoving) |
| Convolutional Neural Network (CNN) | Num-filters | Number of filters in convolution layers | [32,64128] |
| | Filter-size | Size of convolution filters | 3*3 |
| | Pool-size | Size of the pooling filters | 2*2 |
| | activation | Activation function | ReLU |
| | Dropout-rate | Dropout rate in dense layers | 0.5 |
| | Num-dense-units | Number of units in dense layers | 512 |
| Feature Fusion Transformation (FFT) | Attention-heads | Number of attention heads | 8 |
| | Fusion-dim | Dimensions of fused feature | 512 |
| | Num-fusion-layers | Number of layers in fusion transformer | 4 |
| | Dropout-rate | Dropout rate in fusion layers | 0.2 |
| Classification layer | activation | Activation function | SoftMax |
| | Num-classes | Number of output classes | 4 |
| Training parameters | Learning-rate | Learning rate for the optimizer | 0.001 |
| | Batch-size | Number of samples per batch | 32 |
| | Epochs | Number of training epochs | 20 |
| | Optimizer | Optimization algorithm | Adam |
| | Loss-function | Loss function | Categories cross-entropy |
| | Evaluation-metrics | Metrics for evaluating the model | Accuracy, precision, recall, f1 score |

Different parameters are chosen and adjusted to achieve the best results in different stages of the pipeline. The parameters are divided into groups according to the components they correspond to ViT, CNN, FFT, classification layer and training parameters. The architecture of the network consists of layers with varying numbers of filters, filter sizes and pooling operations. The activation function used is ReLU and dropout is applied in the dense layer to avoid overfitting during model training. The transformer employes several attention heads to focus on diverse aspects of the input, capturing more information from the data. The fusion dimensions and the number of layers in the transformer are chosen to ensure that the fused features are stable and semantically significant. The feature fusion network is designed to process the fused features extracted from the transformer in greater depth. It includes multiple fully connected layers with carefully connected units, activation functions and dropout rates to enhance the learned features before classification. The classification layer is the final step in the model, where refined features are classified into one of the predefined classes: far, close, moving, or fast-moving. The *SoftMax* activation function produces the probabilities of each class and makes the final prediction. The values of all the parameters are illustrated in Table 2.

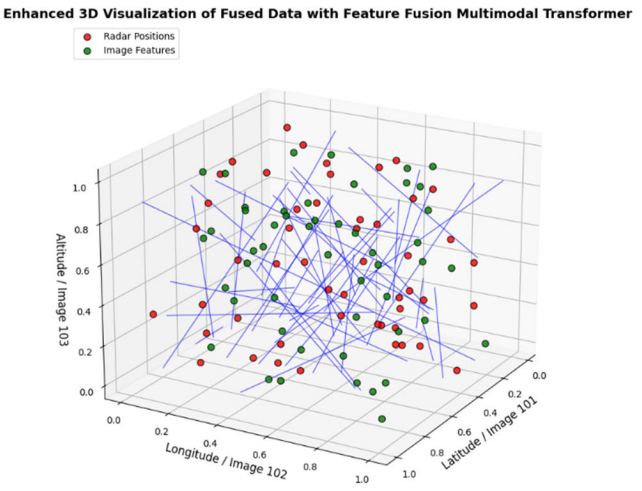


**FIGURE 4.** Fusion results after applying FFMT.

Figure 4 provides a comprehensive visualization of the fusion process used in the model to integrate radar and image data for enhanced object detection and classification. The 3D plot shown in the figure illustrates how the feature fusion method integrates the information from the two data sources, demonstrating the model's ability to handle and integrate multimodal inputs.

The Radar positions are shown as red dots presenting coordinates obtained from the radar sensor. These positions include distance, speed, angle and objects, which constitute important real-time spatial information. The image features are depicted as green dots, representing the feature vectors of the images obtained from the ViT model. The fusion step performed by the FFMT enables the model to create a fused representation of the scene from the radar and image features. The alignment of radar and image features within the fused space indicates the effectiveness of the multimodal fusion process. This fusion demonstrates the model's capability to leverage both data types to improve object identification and classification.

Therefore, this step of the model plays a key role in enhancing object localization and classification enabling efficient and rapid detection of objects. Figure 4 emphasizes the model's ability to integrate and use data from various sources to enhance the spatial orientation and decision-making, which is essential for navigation assistance for the visually impaired. The fusion of radar and image data results in a stable and enhanced perception of the environment, allowing the model to accurately classify objects as close, far, moving or fast-moving, allowing the model to accurately classify objects as close, far, moving or fast-moving, thereby improving navigation safety in real-time. The FFMT ensures that the fused data offers an integrated view to the model, enabling it to classify and detect objects with high accuracy. This integration results in a reliable and responsive navigation system capable of providing real-time assistance to visually impaired users.

According to Table 3, various fusion approaches have been employed across different modalities and sensors. The fusion models are compared with the baseline models that were built

**TABLE 3.** Performance comparison of various multimodal techniques of existing work with fusionsight model.

| Model | Fusion Type | Dataset type/ Sensors | Performance |
|---|---|---|---|
| Multi-modal fusion transformer [26] | ViT, BERT | Image, Text | Remarkable, but lacks real-time performance validation. |
| Multi-modal fusion Transformer [27] | ViT, BERT | Image, Text | 93% accuracy in image-text classification tasks. |
| Multi-modal fusion Transformer [28] | CNN, ViT | Image, Numeric | Superior performance in combining structured and unstructured data. |
| Multi-modal fusion Transformer [29] | CNN, ViT | Camera, Radar, LiDAR | Remarkable, but computationally intensive. |
| EfficientQ3M [30] | Modality Transformer | LiDAR, Sensor | Highly efficient, optimized for energy usage. |
| Multi-modal fusion network[31] | discrete wavelet Transformer | Radar, Camera | More efficient and effective for low-cost applications. |
| Multi-modal Transformer[32] | Transformer model | Image, Sensor data and Text | Impressive performance across diverse modalities. |
| FusionSight model (proposed model) | Vit, CNN | Image, Radar | Improved to 99% accuracy, with faster computational time and reduced resources uage compared to state-of-the-art methods. |

using images and radar or LiDAR. The results indicate that the proposed fusion of the two modalities provides better classification and accuracy than the individual modalities. Therefore, the efficiency of the proposed multimodal fusion is demonstrated. Therefore, the efficiency of the proposed multimodal fusion is demonstrated.

Both approaches have shown effectiveness in some aspects; however, the proposed model includes several novelties that make it distinct from previous work. The FusionSight model employs cross-attention to identify the most relevant

features of the two modalities. This mechanism enables the model to achieve higher accuracy in integrating data from diverse sources, such as images and radar signals, which significantly improves its overall performance.

In contrast to the previous fusion approaches that are inclined to favor some of the modalities over others, the proposed model is fair in that it does not allow any of the modalities to overpower the others during the fusion process. This balanced fusion is crucial in scenarios where object classification and detection accuracy depend on combining information from different sensors. The model achieved an overall accuracy of 99% with individual class accuracies fast-moving 99.0%, close 98.50%, far 99.00% and moving 99.02%.

Thresholds set for labeling.
If volume $< -0.5$:
    Labels. Append (0) # 'close'
Elif $-0.5$ <= volume $< 0.5$:
    Labels. Append (1) # 'far'
Elif 0.5 <= volume $< 1.5$:
    Labels. Append (2) # 'moving'
Else:
    Labels. Append (3) # 'fast-moving'
Return labels.

The high accuracy across all the classes shows that model can distinguish between the different obstacle classes hence making it very reliable when it comes to real-time applications (see Figure. 5). The precision, recall and F1 scores were high across all the trails, which indicates that the model was able to identify obstacles with high accuracy and low false positive and false negative rate. Mathematical equations are used to measure accuracy, precision, recall and F1 score, which are key metrics in assessing the quality of classification model.

Accuracy measures the proportion of correctly classified instances out of the total instances. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

Precision measures the proportion of true positive and predictions among all positive predictions made by the model.

$$Precision = \frac{TP}{TP} + FP, \tag{12}$$

Recall measures the proportion of actual positives that were correctly identified by the model.

$$Recall = \frac{TP}{TP} + FN, \tag{13}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{14}$$

Figure 5 illustrates the model's accuracy progression across 20 epochs for both the training and real-time datasets. The training accuracy steadily increases from an initial value of approximately 40% to over 90%, demonstrating the model's ability to effectively learn from the data. Similarly, the real-time dataset shows a consistent improvement in accuracy, achieving over 98% by the 20$^{th}$ epoch by using real-time

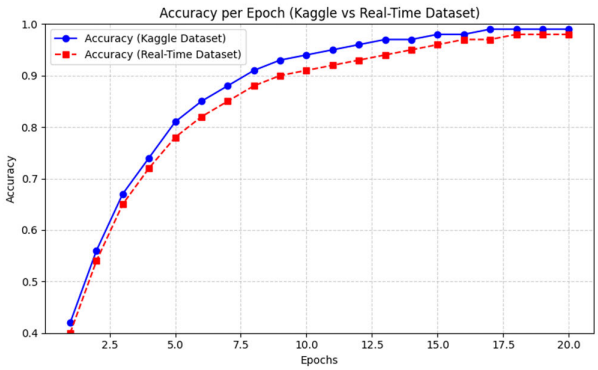


**FIGURE 5.** Classification Accuracy of the FusionSight model.

dataset and 99% accuracy achieved by using static dataset. The comparable accuracy trends for both datasets highlight the model's robustness and generalization capabilities, even when applied to real-world scenarios. The slight difference between the training and real-time accuracy emphasizes the practical challenges in real-time obstacle detection and classification.

These results confirm the proposed FusionSight model's feasibility for use in both controlled and real-time settings, ensuring its reliability for visually impaired individuals.

Figure 7 illustrates the model's loss progression over 20 epochs for both the training and real-time datasets. The sharp decline in loss value from the 1$^{st}$ epochs to the 10$^{th}$ epoch highlights the model's ability to fine-tune its parameters and reduce errors effectively, indicating a strong learning curve. The training loss decreases steadily, demonstrating that the model successfully minimizes the difference between the predicted and actual values.

For real-time dataset, the loss also exhibits a significant drop during the initial epochs and stabilizes at a minimal value by the 20$^{th}$ epoch. This consistent performance across both datasets underscores the model's capability to adapt to varying conditions and maintain low error rates in real-world scenarios.

The results demonstrate the FusionSight model's learning efficiency and reliability, ensuring it can provide accurate predictions in both controlled and real-time environments, which is essential for its practical application in obstacle detection for visually impaired individuals.

The F1-score that considers both precision and recall, reflects the changes in both and increases from 41% to 98%. This implies that as the model increases in its precision and recall, the overall performance of the model in balancing these metrics also increases. This indicates that the model has reached an optimal state where further training will not significantly improve accuracy and other performance indicators. This is a very important observation when deciding on the number of epochs to train the model in future iteration of the model. Figure 8 shows the ROC curve of the FusionSight model's performance of real-time dataset across four object classes: close, far, moving and Fast-moving. Each curve represents the trade-off between the True Positive

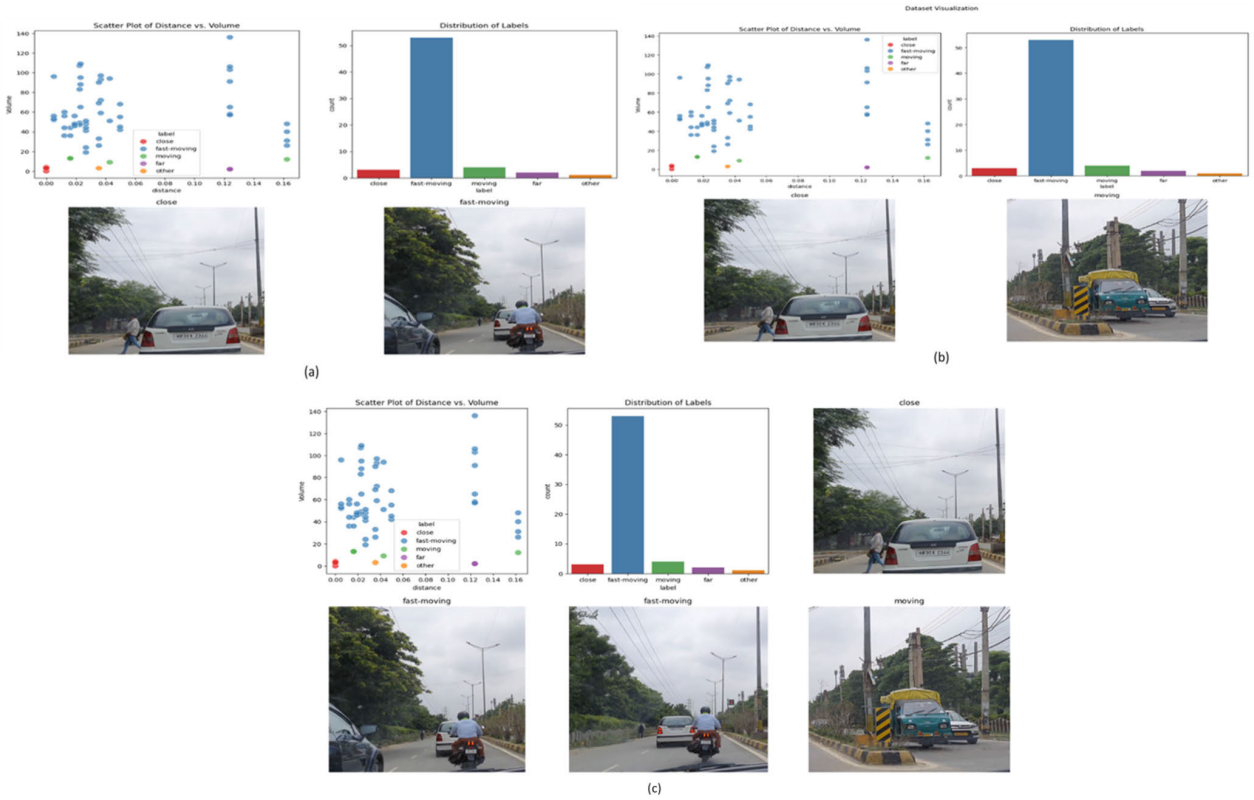


**FIGURE 6.** Qualitative output of object detection and classification with scatter plots of the objects with labels and number of assurance of classes, (a) close, fast-moving, (b) close, moving, (c) close, fast-moving, fast-moving, moving.

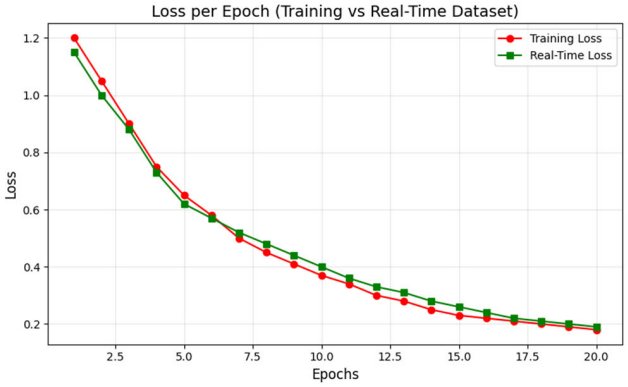


**FIGURE 7.** Classification Loss Results of the FusionSight model.

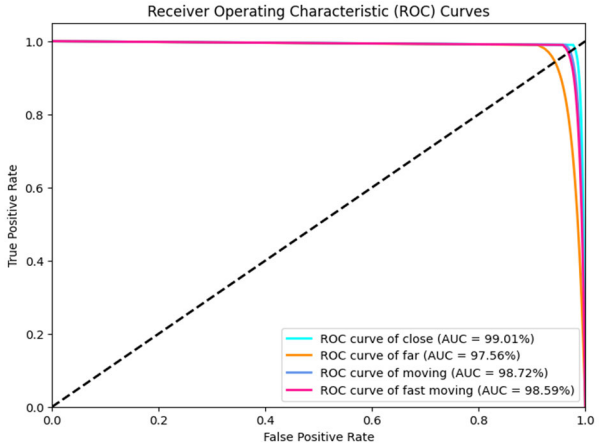


**FIGURE 8.** ROC curve for each class of FusionSight model.

Rate (TPR) and False positive Rate (FPR) for the respective class. The area under the curve demonstrates the model's high classification performance. The ROC curves confirm the robustness and precision of the model, with higher AUC values indicating excellent discriminatory power across all object categories. The dashed diagonal line represents a random classifier's performance (AUC=50%) serving as a baseline for comparison.

The confusion matrix was utilized to evaluate the effectiveness of the FusionSight model, which classifies objects into four categories: ''close, far, moving and fast moving'' using real-time dataset as shown in Figure 9. The rows of the matrix represent the actual classes of the test samples (truth classes), while the columns indicate the predicted classes made by the model. Diagonal values in the confusion matrix show the number of samples that were correctly classified for each class, and higher values on the diagonal indicate strong model performance.

Upon analyzing the confusion matrix, we observe that many samples were correctly classified, as indicated by the high values on the diagonal. This reflects the model's strong performance in distinguishing between the four objects categories, achieving very high classification accuracy. However, there are a few values off the diagonal, representing misclassified samples. These misclassifications are minimal, demonstrating that the model is nearly 100% correct with

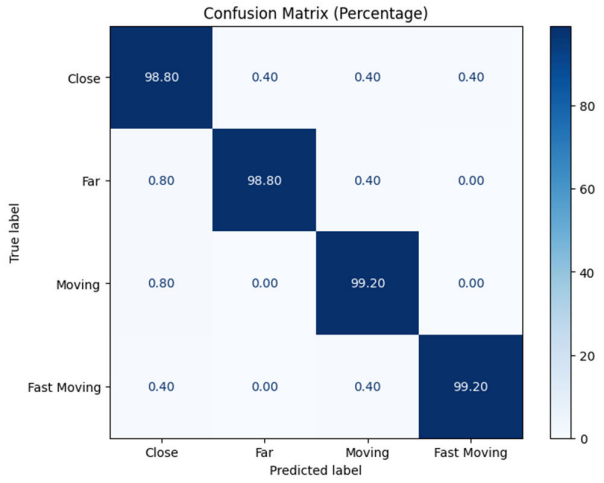


**FIGURE 9.** Confusion matrix of the accuracies achieved by FusionSight model.

a high recall rate. The reported accuracy of 99% considers validation over multiple evaluation folds to ensure robustness, which is slightly less than the near-100% accuracy seen in the confusion matrix due to variations in validation data splits or random initialization factors during training. The few misclassifications, although negligible, indicate potential areas for improvement, such as differentiating between moving and fast-moving objects. The confusion matrix thus confirms that the model is highly effective with only minor areas requiring further fine-tuning to push the accuracy even closer to 100%.

## V. DISCUSSION

The FusionSight model represents a groundbreaking approach to object classification by integrating radar and image data through multimodal fusion. While multimodal fusion has been explored in previous studies, FusionSight uniquely combines ViT, CNN and FFMT to achieve exceptional accuracy, leveraging complementary sensor data. This enables the model to address challenges such as overlapping objects, occlusions and dynamic lighting, which are often problematic for single-modality methods.

As Compared to the baseline models, FusionSight consistently outperforms in real-world scenarios. For instance, in low-light environments, where image-only methods struggle, FusionSight achieves reliable detection by incorporating radar features. Similarly, in high-speed or occlusion-heavy scenarios, the fusion mechanism enhances classification reliability, ensuring practical usability for visually impaired users.

The high-level efficiency of the FusionSight model is reflected in its impressive accuracy, achieving 99% accuracy in object classification. This is supported by metrics such as precision, recall and F1 score, all of which demonstrate the model's proficiency in correctly classifying objects across four distinct classes. The detailed confusion matrix analysis further shows the ability of the model to accurately classify objects into the four defined classes namely ''close, far. Moving. fastmoving'' with minimal misclassification.

Moreover, when evaluated on real-time datasets, the model consistently achieved comparable results demonstrating its ability to generalize to dynamic environments. For real-time applications, the FusionSight model shows reliable detection and classification accuracy even under varying conditions, such as outdoor traffic scenarios involving complex obstacles and diverse lighting. These findings validate the model's practicality and performance in real-world settings.

To further validate these results, we conducted an ablation study. In this study, we systematically evaluated the performance of the model by isolating key components, such as radar-only data, image-only data and removing the Feature Fusion Transformer. The results highlighted that the full multimodal configuration significantly outperforms partial configurations, confirming the importance of each component in achieving optimal performance. The steady increases in accuracy, as evident from the epochs during training, demonstrate the model's high learning capacity and its ability to generalize new data.

However, it is important to note that the proposed model achieves high accuracy on the given datasets, it may require further adjustments or fine-tuning in other scenarios or with different sensors. Factors such as sensor quality, variability of the data, and environmental conditions may affect performance. Future work could focus on improving the model's adaptability and robustness in more diverse conditions.

Another strength of this study is the demonstration of the use of the model in a cloud-connected mobile system developed for the visually impaired. This implementation showcases how the model can be deployed in practice and used to build assistive technologies that operate in real-time and in real-world conditions. The model processes radar and image data on the cloud, converting it into audible outputs, making it actionable for visually impaired users navigating independently through complex environments.

Furthermore, the modularity of the model's architecture makes it possible to apply the model in other domains that require multimodal data fusion. They could include self-driving cars, robots and security systems where the fusion of multiple-sensor data is important for making the right decisions.

## VI. CONCLUSION

The proposed FusionSight model shows promising accuracy and flexibility that can be implemented at a large scale using inexpensive solutions to make it available to visually impaired people in various contexts. Due to the attention mechanism and dense layers in the Feature Fusion Network, we ensure that the features from both radar and image are the ones relevant for an object classification making it more accurate. The experimental results obtained prove the efficiency of the proposed approach and show high accuracy, precision, recall and F1-score. The confusion matrix analysis and epoch-wise performance visualization also provide evidence that the proposed model is more accurate and generalized than the existing methods. In addition to achieving new state-of-the-art performance in multimodal fusion for

object detection and classification, this work also opens new avenues for navigation for real-time applications in dynamic scenarios. The extension of this model into a cloud-connected mobile application also demonstrates the practicality of this model for real-world applications, particularly in assistive technology for the visually impaired. Therefore, this work proves that it is possible to use multimodal deep learning for object detection and recognition which opens the path to the development of new approaches to sensor fusion and real-time classification systems.

## VII. LIMITATIONS AND FUTURE DIRECTIONS

However, the FusionSight model has its merits and it is still possible to identify some of the improvements that can be made. The model has a high computational complexity due to the application of transformers and deep networks, which may be problematic in the case of limited resources. Future research could be directed towards the fine-tuning of the architecture of the model to make it less computationally intensive while at the same time improving its performance. Another limitation of the model is its reliance on Kaggle datasets, which may not fully represent the variability and complexity of real-world environments. For instance, curated datasets often lack edge cases such as extreme weather conditions, diverse urban settings, and sensor malfunctions, which can significantly impact model performance. This reliance restricts the generalizability of the model, highlighting the need for testing on real-world datasets collected in dynamic, uncontrolled condition.

However, the model's dependencies on certain types of sensors and data might reduce its flexibility, and future work could investigate how other data types including LiDAR or sound, can be integrated to expand the model's utility.

Another direction for further research is the exploration of more complex and adaptive fusion methods to enhance the model's performance during the data fusion process. For instance. Approaches like meta-learning reinforcement learning could be investigated to enable the model to dynamically learn and adapt the fusion strategy based on the context or the nature of the incoming data. This adaptability could significantly improve the model's robustness in diverse and unpredictable real-world scenarios. Additionally, integrating advanced attention mechanism or neural architecture search (NAS) techniques could further optimize the fusion process, enabling the model to weigh features from radar and image data more effectively.

The FusionSight multimodal fusion model represents a theoretical and practical advancement in the field of object classification. Its unique architecture, combining transformer-based feature extraction and a feature fusion network, enables high accuracy and robust performance. Despite its reliance on curated datasets, its modular design and flexibility make it a promising solution for various real-world tasks. This work not only demonstrates the potential of multimodal deep learning but also sets the foundation for further innovations in sensor data fusion and real-time classification systems.

## CONFLICTS OF INTEREST

The authors have disclosed no potential conflicts of interest concerning the authorship and/or publication of this article.

## AUTHORS' CONTRIBUTIONS

Sunnia Ikram contributed to the methodology, data collection and data coding, data validity, and writing. Imran Sarwar Bajwa contributed to the data analysis and supervision. Amna Ikram contributed to the data collection. Haleema helped in sensor connectivity and data collection.

Co-author M. Abdullah-Al-Wadud contributed to the final revision of the manuscript by providing valuable insights into the model improvements and experimental validation. Their input led to refinements in the discussion of results, particularly with respect to performance analysis and comparison with existing methods.

## FUNDING AND ACKNOWLEDGMENTS

No AI tool has been used to prepare this manuscript.

All the figures in the article were created by the author. Can be submitted in an editable format on request. The authors acknowledge the Researchers Supporting Project number (RSPD2025R951), King Saud University, Riyadh, Saudi Arabia.

## REFERENCES

[1] C. Ramisetti, T. Neeraj, P. Surya, G. M. Kumar, N. A. Vignesh, A. K. Panigrahy, A. M. V. Bharathy, and N. Kumareshan, "An ultrasonic sensor-based blind stick analysis with instant accident alert for blind people," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2022, pp. 1–13.

[2] D. Nunes, J. Fortuna, B. Damas, and R. Ventura, "Real-time vision based obstacle detection in maritime environments," in *Proc. IEEE Int. Conf. Auto. Robot Syst. Competitions (ICARSC)*, Apr. 2022, pp. 243–248.

[3] E. H. Assaf, C. von Einem, C. Cadena, R. Siegwart, and F. Tschopp, "High-precision low-cost gimballing platform for long-range railway obstacle detection," *Sensors*, vol. 22, no. 2, p. 474, Jan. 2022.

[4] D. He, "Urban rail transit obstacle detection based on Improved R-CNN," *Measurement*, vol. 196, Jun. 2022, Art. no. 111277.

[5] L. Guan, L. Jia, Z. Xie, and C. Yin, "A lightweight framework for obstacle detection in the railway image based on fast region proposal and improved YOLO-tiny network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–16, 2022.

[6] S. Ikram et al., "A IoT-enabled obstacle detection and recognition technique for blind persons," Islamia Univ. Bahawalpur, Pakistan, 2024.

[7] A. K. Sahoo and S. K. Udgata, "Material classification based on non-contact ultrasonic echo signal using deep learning approach," *Proc. Comput. Sci.*, vol. 235, pp. 606–616, Jan. 2024.

[8] S. Khan, H. Ali, Z. Ullah, and M. F. Bulbul, "An intelligent monitoring system of vehicles on highway traffic," in *Proc. 12th Int. Conf. Open Source Syst. Technol. (ICOSST)*, Dec. 2018, pp. 71–75.

[9] N. A. Jasman, M. F. I. M. Jalil, A. Mukhtar, K. S. M. Sahari, and M. E. Rusli, "IoT-based obstacle detection system for visually impaired person with smartphone module," *J. Adv. Inf. Technol.*, vol. 13, no. 4, pp. 368–373, 2022.

[10] M. Skoczeń, M. Ochman, K. Spyra, M. Nikodem, D. Krata, M. Panek, and A. Pawłowski, "Obstacle detection system for agricultural mobile robot application using RGB-D cameras," *Sensors*, vol. 21, no. 16, p. 5292, Aug. 2021.

[11] M. S. Latif, "Pest prediction in Rice using IoT and feed forward neural network," *KSII Trans. Internet Inf. Syst.*, vol. 16, no. 1, pp. 133–152, 2022.

[12] N. Rachburee and W. Punlumjeak, "An assistive model of obstacle detection based on deep learning: YOLOv3 for visually impaired people," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 11, no. 4, p. 3434, Aug. 2021.

[13] M. Zou, J. Yu, Y. Lv, B. Lu, W. Chi, and L. Sun, "A novel day-to-night obstacle detection method for excavators based on image enhancement and multisensor fusion," *IEEE Sensors J.*, vol. 23, no. 10, pp. 10825–10835, May 2023.

[14] J. Qin, R. Sun, K. Zhou, Y. Xu, B. Lin, L. Yang, Z. Chen, L. Wen, and C. Wu, "LiDAR-based 3D obstacle detection using focal voxel R-CNN for farmland environment," *Agronomy*, vol. 13, no. 3, p. 650, Feb. 2023.
[15] G. Zhu, "High precision rail surface obstacle detection algorithm based on 3D imaging LiDAR," *Opt. Lasers Eng.*, vol. 178, Jul. 2024, Art. no. 108206.
[16] R. Păpară, L. Grec, I.-A. Potarniche, and R. G. Voichita, "Testing of indoor obstacle-detection prototypes designed for visually impaired persons," *Appl. Sci.*, vol. 14, no. 5, p. 1767, Feb. 2024.
[17] T. Kocejko, N. Matuszkiewicz, P. Durawa, A. Madajczak, and J. Kwiatkowski, "How integration of a brain-machine interface and obstacle detection system can improve wheelchair control via movement imagery," *Sensors*, vol. 24, no. 3, p. 918, Jan. 2024.
[18] L. Bougheloum, M. B. Salah, and M. Bettayeb, "Outdoor obstacle detection for visually impaired using AI technique," in *Proc. ASU Int. Conf. Emerg. Technol. Sustainability Intell. Syst. (ICETSIS)*, Jan. 2024, pp. 628–633.
[19] H. G. Kusneniwar, S. Ghosh, and S. Sen, "SonicGlass: An obstacle detection and navigation system using smartglass-based ultrasonic sensors," in *Proc. 16th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2024, pp. 603–607.
[20] J. Deepa, P. M. Adeline, S. S. S. Madhumita, and N. Pavalaselvi, "Obstacle detection and navigation for the visually impaired," in *Proc. 9th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2023, pp. 905–909.
[21] S. Shahani and N. Gupta, "The methods of visually impaired navigating and obstacle avoidance," in *Proc. Int. Conf. Appl. Intell. Sustain. Comput. (ICAISC)*, Jun. 2023, pp. 1–6.
[22] A. Ikram, W. Aslam, R. H. H. Aziz, F. Noor, G. A. Mallah, S. Ikram, M. S. Ahmad, A. M. Abdullah, and I. Ullah, "Crop yield maximization using an IoT-based smart decision," *J. Sensors*, vol. 2022, pp. 1–15, May 2022.
[23] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7073–7083.
[24] R. Shi, S. Yang, Y. Chen, R. Wang, M. Zhang, J. Lu, and Y. Cao, "CNN-transformer for visual-tactile fusion applied in road recognition of autonomous vehicles," *Pattern Recognit. Lett.*, vol. 166, pp. 200–208, Feb. 2023.
[25] A. Naseer and A. Jalal, "Multimodal objects categorization by fusing GMM and multi-layer perceptron," in *Proc. 5th Int. Conf. Advancements Comput. Sci. (ICACS)*, Feb. 2024, pp. 1–7.
[26] A. Azeem, Z. Li, A. Siddique, Y. Zhang, and S. Zhou, "Unified multi-modal fusion transformer for few shot object detection for remote sensing images," *Inf. Fusion*, vol. 111, Nov. 2024, Art. no. 102508.
[27] A. Xiang, Z. Qi, H. Wang, Q. Yang, and D. Ma, "A multimodal fusion network for Student emotion recognition based on transformer and tensor product," 2024, *arXiv:2403.08511*.
[28] X. Ma et al., "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 17, pp. 20116–20128, 2024.
[29] S. Y. Alaba, A. C. Gurbuz, and J. E. Ball, "Emerging trends in autonomous vehicle perception: Multimodal fusion for 3D object detection," *World Electr. Vehicle J.*, vol. 15, no. 1, p. 20, Jan. 2024.
[30] M. R. Van Geerenstein, F. Ruppel, K. Dietmayer, and D. M. Gavrila, "Multimodal object query initialization for 3D object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 12484–12491.
[31] S. Y. Alaba and J. E. Ball, "Transformer-based optimized multimodal fusion for 3D object detection in autonomous driving," *IEEE Access*, vol. 12, pp. 50165–50176, 2024.
[32] Y. Lu, X. Lu, L. Zheng, M. Sun, S. Chen, B. Chen, T. Wang, J. Yang, and C. Lv, "Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems," *Plants*, vol. 13, no. 7, p. 972, Mar. 2024.

**SUNNIA IKRAM** received the M.S. degree in computer science from The University of Lahore. She has made significant contributions to the field of IoT, artificial intelligence, and sensor fusion technology. She is a Distinguished Academic and a Researcher specializing in computer networks, the Internet of Things (IoT) and deep learning applications. She is currently serving as a Lecturer with the Software Engineering Department, The Islamia University of Bahawalpur, Pakistan. As a dedicated educator and researcher, she continues to push the boundaries of AI-driven solutions, integrating computer vision, pattern recognition, and real-time sensor data processing into practical applications.

**IMRAN SARWAR BAJWA** received the Ph.D. degree in computer science from the University of Birmingham, U.K. He has more than 19 years of teaching and research experience in various universities of Pakistan, Portugal, and U.K. He is currently a Full Professor with the Department of Computer Science and Information Technology, The Islamia University of Bahawalpur. He is the author/editor of 12 books published by IEEE, Springer, and IGI Global. He has more than 200 articles and 3100 citations in Google Scholar. In addition, he has more than 138 articles in Scopus and 1700 citations of the work in Scopus. His Google H-index is 30 and Scopus H-index is 22. His personal impact factor is more than 150. His current research interests include intelligent systems, the IoT, and data analytics. He has been an associate editor and a guest editor of various IEEE and Elsevier journals.

**AMNA IKRAM** received the B.S. and M.S. degrees in computer science from The Islamia University of Bahawalpur, Pakistan, in 2006 and 2016, respectively, where she is currently pursuing the Ph.D. degree in computer science. She is currently a Senior Lecturer with the Department of Computer Science and Information Technology, The Government Sadiq College Women University, Bahawalpur, Pakistan. Her research interest includes cropping yield maximization using AI-driven decision-based methodologies.

**M. ABDULLAH-AL-WADUD** (Member, IEEE) received the B.S. degree in computer science and the M.S. degree in computer science and engineering from the University of Dhaka, Bangladesh, in 2003 and 2004, respectively, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2009. He is working as an Associate Professor with the Department of Software Engineering, King Saud University, Saudi Arabia. Afterwards, he served as a member of the faculty of the Department of Industrial and Management Engineering, Hankuk University of Foreign Studies, South Korea, from 2009 to 2014. He also served as a Lecturer with Daffodil International University, Bangladesh, and East West University, Bangladesh. His research interests include artificial intelligence, optimization, computer vision, cloud computing, recommender systems, software engineering, sensor, and ad hoc networks.

**HALEEMA PK** is worked as a Senior Lecturer at the School of Computing and Engineering, University of West London – RAK Branch Campus, UAE. Currently, she is an Assistant Professor at the Rochester Institute of Technology, USA. She has published extensively in international peer-reviewed journals and conferences and actively collaborates on interdisciplinary research projects. Her research interests include machine learning, artificial intelligence, computer vision, and data science applications. She has supervised multiple postgraduate students and contributed to curriculum development in advanced computing disciplines. With a strong background in both academia and industry, she is dedicated to fostering innovation and research excellence.