IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 June 2022, accepted 16 July 2022, date of publication 21 July 2022, date of current version 28 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193097

APPLIED RESEARCH

IBGS: A Wearable Smart System to Assist Visually Challenged

KUN XIA, XUEYONG LI, HAIYANG LIU, MINGLI ZHOU, AND KEXIN ZHU

Department of Electrical Engineering, University of Shanghai for Science and Technology, Yangpu, Shanghai 200093, China

Corresponding author: Xueyong Li (llxxyyexcellent@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51207091, and in part by the Key Project of National Natural Science Foundation of China under Grant 51637001.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Experimental Committee of the University of Shanghai for Science and Technology under Application No. 10-21-347-001, Measures for Ethical Review of Biomedical Research Involving Humans.

ABSTRACT Traditional blind guide devices are expensive and large. In this study, an intelligent blind guide system (IBGS) was introduced. GD32 is used as the main control chip, it cooperates with various functional modules to realize traffic light recognition, obstacle avoidance, payment, and navigation functions on the basis of speech recognition. At the same time, IBGS uses WIFI instead of Bluetooth to get rid of the dependence on smart phones. In addition, a cloud database was built, and the Internet of Things technology was used to realize the information interconnection between the IBGS, database, and the guardian's mobile terminal. In order to better realize the function of speech recognition, this paper proposes a Conv-Transformer Transducer (ConvT-T) speech recognition framework based on Weak-Attention Suppression (WAS), which improves the efficiency of multi-head attention through WAS. The proposed method achieved a word error rate (WER) of 3.2% in test-clean and 7.9% in test-other on the LibriSpeech ASR corpus with only 73M parameters, which reduces the WER by 0.3% and 0.4% compared to ConvT-T respectively, indicating that WAS can effectively play the role of suppressing non-critical attention on the ConvT-T framework. At the same time, the IBGS was tested in a comprehensive application scenario. In the outdoor traffic light recognition, speech recognition, and obstacle avoidance tests, the accuracy rates were 92.33%, 90.33%, and 96.67% respectively. The results show that the IBGS can effectively deal with various challenges encountered by the blind during their daily outdoor walking.

INDEX TERMS Visually impaired, wearable device, traffic light recognition, speech recognition.

I. INTRODUCTION

Due to their limitations of vision, blind people have many inconveniences in work, study, and even daily outdoor walking. According to the World Health Organization, at least 2.2 billion people worldwide were visually impaired or blind by 2020.

In recent years, the visually challenge to blind people has received more and more attention. With the development of electronic technology in the medical field, smart portable devices are regarded as a new means to solve this problem. Many current outdoor walking solutions for the blind take the form of wearable devices, such as white cane, glasses, gloves, and shoes. These devices are dedicated to solving problems such as obstacle avoidance, object recognition, navigation or calling for help encountered by the blind while outdoor walking. But existing devices often only perform one or two of these functions, and blind people still cannot go out independently with them alone. Some devices have more comprehensive functions, but at the same time, they also bring problems of high cost or cumbersome wearing that cannot be ignored. There are also some studies that have to use the powerful functions of smartphones in order to complete more comprehensive services, but smartphones are unnecessary for blind people because they cannot use it as easily as ordinary people.

The associate editor coordinating the review of this manuscript and approving it for publication was Chan Hwang See.

Based on the collaborative work between GD32 and multi-modules, this paper proposes an intelligent blind guide system (IBGS) that helps blind people to go out independently. The system is realized in the form of glasses clip, which can be clipped on the sunglasses that blind people wear daily, which solves the problems of the thick and single size of the existing smart glasses. The glasses clip adopts a sound-absorbing cavity structure, which effectively blocks the sound interference between the microphone and the speaker, and improves the efficiency of the speech interaction function. On the basis of speech interaction, IBGS realizes the function of traffic light recognition based on YOLO [1] framework and OpenCV [2] vision library. In addition, IBGS uses laser ranging to help the blind avoid obstacles. Furthermore, IBGS is equipped with a Near Field Communication (NFC) chip [3], the NFC-related support driver is added to the Android source code, and the corresponding baseband circuit and radio frequency circuit are matched to realize the mobile payment function. In order to solve the situation that many blind solutions have to be coupled to the smartphone, IBGS applies the OneNET cloud database, conducts WIFI wireless communication through the MQTT protocol, and transmits the longitude and latitude information obtained by the built-in GPS of the glasses clip in real time, helping the blind to go out independently without smartphone, and facilitating the guardian to obtain the wearer's movement track in time. At the same time, the Internet of Things (IoT) technology is used to connect the system, database, and guardian's mobile terminal to form a complete closed-loop feedback mechanism.

The contributions of this paper are as follows:

This paper proposes IBGS, which can detect traffic lights in real time, effectively avoid obstacles, locate without smartphones, connect with guardians, and achieve the above functions through speech interaction.

On the basis of the Conv-Transformer Transducer (ConvT-T) speech recognition framework, WAS learning is added to improve the efficiency of multi-head attention, thus effectively reduces word error rate (WER).

The YOLOv3-tiny network is embedded into mobile devices and successfully applied to traffic light recognition of daily life.

The NFC function is embedded in the IBGS, which solves the payment difficulties encountered by the blind in daily life.

The IoT technology is used to connect multiple platforms and realize the information docking between embedded devices, mobile phones, and cloud databases.

A comprehensive application scenario test is carried out to prove that IBGS can effectively solve the problems of independent outdoor walking for the blind.

The rest of this paper is arranged in the following manner. Section II presents recent advances in current research on assisted mobility for the blind. Section III describes the system. Section IV introduces the two main algorithms involved in this system, i.e., the collected data processing methods. Section V verifies the performance of the proposed system through experiments in real scenarios. Section VI summarizes the current research progress and provides an outlook for the future.

II. RELATED RESEARCH

The current research focuses on solving several main problems encountered by blind people in outdoor walking: positioning, image recognition, obstacle avoidance, and help calling [4]–[8]. Masud *et al.* proposed an intelligent white cane, which can help blind people avoid obstacles and simultaneously announce the category of obstacles by speech feedback [4]. The white cane proposed by Dharan *et al.* can detect stairs or pedestrians ahead, while flame and water sensors are attached to monitor the happenings of the surroundings to protect the user. The research combines various sensors with Google's speech assistant model to provide users with vibration or speech feedback. Also, the white cane is equipped with GPS sensor to provide location information to registered mobile numbers [5]. Nandalal *et al.* studied an intelligent shoe that uses ultrasonic sensors to complete obstacle detection. It is worth mentioning that in this study, the distances of obstacles were divided into three cases, corresponding to the frequencies of three vibration reminders respectively. In addition, they implement the SOS function by sending an alarm text message to the connected mobile phone via the GSM module [6]. The smart shoe proposed by Singh *et al.* can not only detect obstacles in front but also detect the presence of any major pits on the way. Also it contains a panic button which can send SMS alerts in the case of emergency [7]. The smart helmet proposed by Huang *et al.* can also realize obstacle avoidance and GPS-based positioning technology. In addition, they realize traffic light recognition through binocular cameras. However, binocular cameras are particularly sensitive to ambient lighting and are not suitable for scenes that lack texture [8].

While meeting the common needs of blind people, many studies have also explored other services to provide more detailed guarantees for blind people to go out [9]–[12]. The tracking system proposed by Ashiq *et al.* provide navigation in real-time using an automated voice. This work develops a web-based application to help guardians track the user's outdoor activities, including obtaining location and snapshots. The snapshots are acquired using the deep Convolution Neural Network (CNN) model, and the dataset contains more than 1,000 categories [9]. Kumar *et al.* proposed a wearable mask and a stick, which can achieve path edge detection applying the YOLOv3 algorithm, whereas the dataset contains only 25 categories. It is worth mentioning that this work uses bone conduction headphones to provide information feedback to the user [10]. The wearable assistive system proposed by Chang *et al.* consists of a pair of sunglasses, a waist-mounted device, and a walking stick. They can realize traffic light and zebra crossing recognition, and additionally add anti-fall function [11]. The smart crutches proposed by Romadhon and Husein use an ultrasonic sensor and a water sensor, and additionally add a heart sensor to detect the user's health status [12].

Smartphones have very powerful features and excellent operating systems, and many researches are devoted to using smartphones to obtain better functions [13]–[19]. Chandna and Singhal proposed a deep learning based outdoor smartphone navigation system approach which will assist the user in crossing the road by detecting the crosswalks from a mobile device. To accomplish this work, they built a YOLOv5 architecture that can be used on a smartphone [13]. References [14]–[17] explores the services provided for the outdoor walking of the blind through the combination of wearable devices and smartphones. While using sensors to detect obstacles, they connect to mobile phones via Bluetooth to enable speech navigation, traffic sign recognition, or communication with guardians. Martínez-Cruz *et al.* proposed an assistance system in the use of public transportation. The proposed system uses Bluetooth Low Energy technology for location and communication purposes, at the same time they developed a mobile application for user-smartphone interaction. The mobile application provides the relevant information to the user employing verbal instructions, e.g., transportation line, destination, next stop name, and current location [18]. Ali A. *et al.* use Google Glass in combination with a smartphone that acts as a visual assistant. The smartphone device is used to process the captured image. This work contains a dataset of 5000 images, which enables excellent scene recognition. It should not be overlooked that the average price of Google Glass is $1500 [19].

There are also many studies trying to reduce costs as much as possible while providing outdoor walking assistance for the blind [20]–[22]. Biswas *et al.* [20], Ashrafuzzaman *et al.* [21] help blind people avoid obstacles by simply adding ultrasonic sensors and water sensors, and use a buzzer to remind users. Both GPS and GSM components are used to locate and send information to relatives. According to the author's offer, the price of these blind sticks can be kept under $30. The smart glasses proposed by Peyal *et al.* adds more functions. Through this smart eyewear, visual input will convert into the audible signal. Optical character recognition is available for recognizing Bangla and English text. Using speech recognition API users may be able to control electrical gadgets and communicate with caregivers through a mobile app. This system is quoted for about $200 [22].

We present a comparison between the performance of the proposed system and recent similar work in Table 1.

Automatic speech recognition (ASR) [23] has always been an important task in the field of artificial intelligence research. The dominance of Gaussian of mixture-hidden Markov model (GMM-HMM) in speech recognition dates back to the 1980s. Subsequently, deep neural network hidden Markov model (DNN-HMM) emerged along with the development of deep learning [24]. The above-mentioned traditional speech recognition systems are very complex, and usually require separate training of an acoustic model, a language model, and a pronunciation model. Therefore, end-to-end speech recognition methods have been explored, enabling direct mapping of sequences of input acoustic features to text using a single sequence-to-sequence model [25]–[27]. A. Graves proposed a "deep" RNN composed of stacked Long Short-Term Memory (LSTM) layers in 2012 and discussed how to train it for next-step prediction, thus enabling sequence generation. Graves demonstrated the ability of LSTM recurrent neural network (RNN) architecture to generate discrete and real-valued sequences with a complex long-term structure using next-step prediction [28]. After Transformer was proposed in the field of natural language processing, Dong *et al.* proposed a no-recurrence sequence-to-sequence model entirely relies on attention mechanism to learn the positional dependencies — Speech-Transformer [29] in 2018 and achieved remarkable success in the field of speech recognition [30], [31]. Since Transformer cannot be directly used in streaming speech recognition scenarios, in the past two years, people have focused on improving the Transformer model to achieve low-latency streaming recognition. In 2019, Facebook proposed Transformer-Transducer [32]. This structure is based on RNN-T. Before sending acoustic features to audio encoder, in order to reduce delay, truncated self-attention is used instead of full-sequence self-attention to limit the context window. In 2020, based on Facebook, Google used Transformer in both audio encoder and label encoder, and limited the context information of label encoder's self-attention to reduce latency [33]. In addition, Huawei proposed the ConvT-T on the basis of the aforementioned, i.e., using a unidirectional Transformer with interleaved convolution for audio encoding, and adding interleaved convolution between Transformer layers to model future contexts [34]. This architecture achieves performance on LibriSpeech ASR corpus [35], with WER of 3.5% in test-clean and 8.3% in test-other, and is more lightweight on the basis of the aforementioned research. The functions proposed in this paper are based on speech recognition, so ConvT-T is very suitable for this system. At the same time, in order to ensure the accuracy requirements, this paper creatively proposes a method that adds Weak-Attention Suppression (WAS) [36] to the attention mechanism to improve the efficiency of multi-head attention in the Transformer.

III. PROPOSED FRAMEWORK

IBGS is an intelligent system that can help blind people to go out independently. Blind people can wear it to consume, cross the road, and walk on the street safely. This section mainly introduces the composition and design of the system. The system is shown in Fig. 1.

A. STRUCTURE DESIGN

In order to ensure the portability of the equipment, the IBGS is implemented in the form of a glasses clip, as shown in Fig. 2. The outer casing of the glasses clip is modeled by

TABLE 1. Comparison of published works.

Ref.	Detection devices	Special features	Interactive mode		Controllers and sensors	Improved algorithm	Cost ($)
			Command	Feedback			
[4]	White cane	Obstacle detection Object recognition	NM	Speech feedback	Raspberry-Pi 4B Camera Ultrasonic sensor	Viola Jones TensorFlow	NM
[5]	White cane Gloves	Obstacle detection Location tracking Stair forecast	Button	Speech feedback Vibrate	Arduino Raspberry Pi Arduino Mega 2560 Ultrasonic sensor Flame sensor Hydro sensor GPS-GSM modules Staircase sensor Obstacle sensor Vibration sensor	NM	NM
[6]	Shoes	Obstacle detection Emergency call	NM	Vibrate	Arduino Nano Ultrasonic sensor Vibrating motor Battery GSM	NM	NM
[7]	Shoes	Obstacle detection Emergency call	Button	Speech feedback	Arduino Ultrasonic sensors IR sensor Water sensor GSM Audio playback board with speaker/headphone	NM	NM
[8]	Helmet	Obstacle detection Navigation Traffic lights recognition Zebra crossing recognition	NM	Speech feedback Vibrate	Binocular camera GPS BLE Speaker	NM	NM
[9]	DSP Headphones Camera	Navigation Send location to guardian	NM	Speech feedback	Raspberry Pi 3B GPS-GSM modules Text to speech converter	Mobile-net	NM
[10]	Wearable mask White cane	Obstacle detection Object recognition Path edge detection	NM	Speech feedback	Raspberry Pi 3B Ultrasonic Sensor RGB Pi-Camera model B of resolution 5 megapixel Bone conduction headphones	YOLOv3	NM
[11]	Glasses White cane Waist-mounted device	Fall detection Front aerial obstacle avoidance Urgent notification Traffic lights recognition Zebra crossing recognition	NM	Speech feedback	Camera ToF laser-ranging Vibration motor 6-axis motion tracking sensor GPS LPWAN	Inception-based deep learning	NM
[12]	White cane	Obstacle detection Location tracking Emergency call Pulse detection	Button	Speech feedback	Ultrasonic Sensor GPS-GSM modules Pulse heart sensor Water sensor	NM	NM
[13]	Smart phone	Zebra crossing recognition	NM	Speech feedback	Android	YOLOv5	NM
[14]	Gloves Smart phone	Direction of oncoming persons recognition	NM	Speech feedback	Raspberry-Pi 4B Android	OpenCV	NM

TABLE 1. ***(Continued.)*** **Comparison of published works.**

		Traffic lights recognition Road sign recognition			3V DC vibration motors		
[15]	White cane Smart phone	Obstacle detection Navigation Emergency call	Button	Speech feedback	ESP32 Mini Android Detection sensor Vibration module	Azure Custom Vision API	NM
[16]	White cane Smart phone	Obstacle detection Location tracking Emergency call	Button	NM	Raspberry Pi 3B Raspberry Pi camera Ultrasonic sensors	NM	NM
[17]	Belt Smart phone	Obstacle detection Object recognition	NM	Speech feedback	Arduino 1sheeld BLE Ultrasonic sensor IR sensor GSM	NM	NM
[18]	Smart phone BLE beacons	Bus system assistance	Button	Speech feedback	Android BLE	NM	NM
[19]	Google glass Smart phone	Scene recognition	Speech command	Speech feedback	Bone conduction speaker	NM	>1500
[20]	White cane	Obstacle detection Send location to guardian	Remote control	Buzzer	Arduino Ultrasonic-Water sensors GPS-GSM modules Remote Buzzer 9V Battery	NM	28
[21]	White cane	Obstacle detection Send location to guardian	Remote control	Buzzer	Arduino UNO GPS-GSM modules Ultrasonic sensor Water sensor Remote server Buzzer 9-Volt battery	NM	27.73
[22]	Glasses	Object recognition Character recognition Smart home appliance control	Speech command	Speech feedback	Raspberry Pi 4B Camera GPS-GSM modules Remote server Earpiece Lithium-ion 4000mAH	ResNet50 neural network	200
Proposal	Glasses clip	Obstacle detection Navigation Traffic lights recognition NFC payment Emergency call Data cloud storage Guardian monitoring	Speech command	Speech feedback	GD32 Maix Nano SNR9813 WIFI NFC GPS Ranging sensor 3.7V Battery	YOLOv3-tiny Conv-Transformer Transducer with WAS	NM

NM = not mentioned

Creo and printed by precision 3D printing. The material is Polyltie 8200, which is an unsaturated polyester resin with a smooth surface and toughness so that the overall structure of the glasses clip has sufficient strength, and can withstand a certain load and impact. Considering that the glasses clip needed to be fixed on the sunglasses, a precision spring clip is designed to ensure its stability.

In addition, in order to improve the accuracy of the speech interaction function, we first set up ventilation holes on the outside of the glasses clip to ensure the heat dissipation of

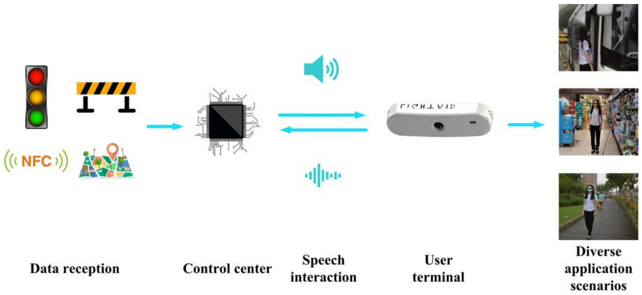
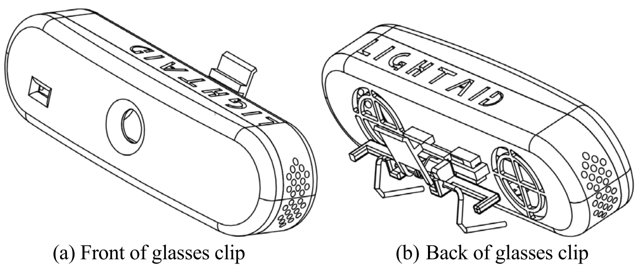


FIGURE 1. Overall system block diagram.



(a) Front of glasses clip (b) Back of glasses clip

FIGURE 2. Model of glasses clip.

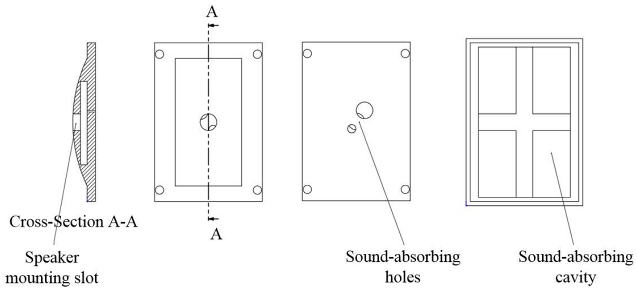


FIGURE 3. Schematic diagram of sound-absorbing cavity.

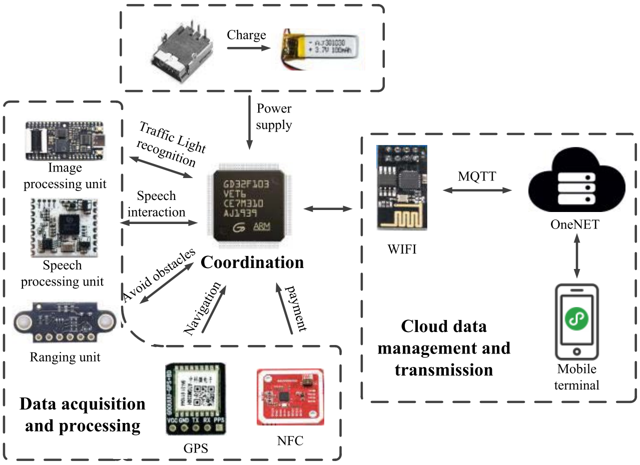
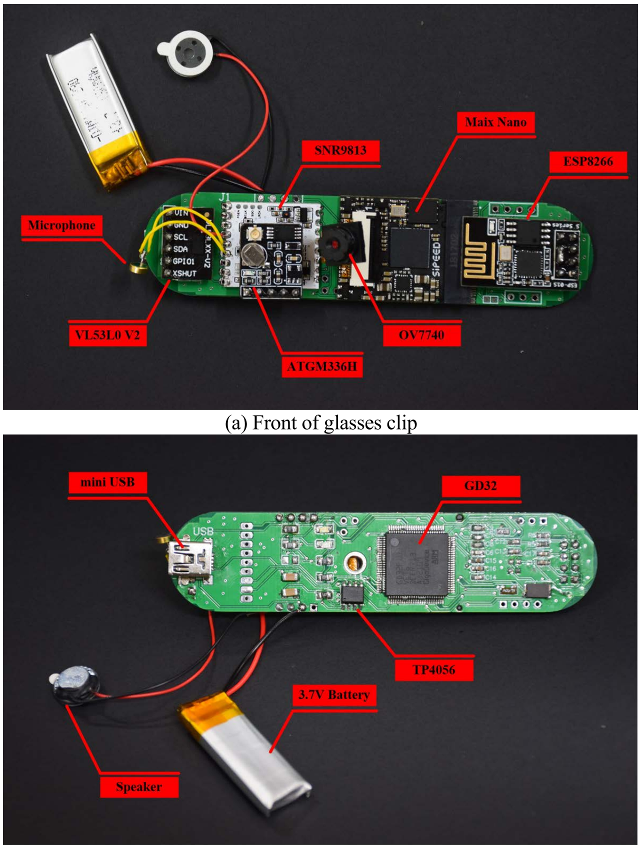


FIGURE 4. Hardware composition of IBGS.



(a) Front of glasses clip

(b) Back of glasses clip

FIGURE 5. Photo of hardware composition. The NFC tag is attached to the shell of the glasses clip

the system. In addition, we set up a sound-absorbing cavity inside to reduce the sound interference between the speaker and the microphone. Fig. 3 shows the structure of the sound-absorbing cavity, i.e., a mounting block is arranged on the microphone, a mounting groove is arranged on the surface of one side of the mounting block, and a loudspeaker is arranged in the mounting groove. A placing cavity for sound-absorbing materials is arranged at the bottom of the mounting groove, and a sound-absorbing hole is arranged on the sidewall of the placing cavity near the loudspeaker for sound waves to enter the placing cavity.

As shown in Fig. 3, when the loudspeaker produces sound waves, the vast majority of the sound waves generated by the loudspeaker can be diffused from the slot of the installation slot to the outside. In addition, a small part of the sound wave is transmitted to the bottom of the mounting groove from the sound-absorbing hole into the placing cavity, thereby being absorbed by the sound-absorbing material in the placing cavity. Compared with installing the speaker in the mounting slot directly, this method can effectively absorb the sound wave transmitted to the bottom of the mounting slot, and reduce the generation of standing waves so that the MCU can receive instructions in less noise.

B. HARDWARE OVERVIEW

As shown in Fig. 4. The IBGS adopts a modular design and is composed of the main control chip, image module, speech module, ranging module, GPS, NFC chip, and WIFI communication unit. The photo of hardware composition is shown in Fig. 5.

MCU adopts GD32F103VET6, which uses ARM® Cortex^{TM}-M3 processor core. Its maximum in-chip Flash is 1024 KB, the RAM up to 256 KB, and the chip power supply voltage ranges from 2.6 V to 3.6 V. It compiles different functions into clear data flow, which is mainly divided into the foreground client side and the background data processing side.

The client side receives the signal through each sensor. The analysis of speech commands is the basis for processing the data of each sensor. The speech processing unit adopts SNR9813 neural network processor, which supports 16 k sampling, 3.3 V power supply, and only 50 mA in standby mode. In this work, we set the corresponding functional commands in advance. When the speech processing unit recognizes a specific command, it will notify the MCU to receive the data from the sensor. Similarly, the speech feedback will also be broadcast to the user through the 0.5 w speaker equipped with the SNR9813.

In the image processing unit, OV7740 is used to collect images, which is small in size, light in weight, it equipped with 2 megapixels, and has a transmission rate of up to 60 fps, whereas the power consumption only needs 60 mA under 3.3 V voltage. Another part of the image processing unit is the Maix Nano, OV7740 transmits the collected data to it for further processing. Maix Nano is a project that transplants Micropython to K210 (a 64-bit dual-core RISC-V CPU with hardware FPU, convolution accelerator, FFT, Sha256), supports regular MCU operations, integrates hardware-accelerated AI machine vision, and is convenient for us to perform model training in this work to meet the needs of IBGS for extremely low cost and small size.

The ranging unit adopts VL53L0 V2 laser ranging chip. It integrates a leading avalanche photon diode array and embeds ST's second-generation FlightSense^{TM}patented technology. The 940nm vertical cavity surface emitting laser of VL53L0 V2 is completely invisible to the human eye, and the built-in physical infrared filter makes it longer ranging, more immune to ambient light, and more stable facing the optical crosstalk of cover sheet. When powered by 3.3 V, it only needs 6 mA in working mode and 5 μA in standby mode.

For the proposed positioning and navigation functions, we choose the ATGM336H module to implement. ATGM336H supports Beidou /GPS /GLONASS satellite system and 3.3 V power supply, the power supply current only needs 20 mA. It has rechargeable electronics onboard to speed up the hot-start star search process. It adopts TTL level UART interface, which is convenient to communicate with GD32. At the same time, it has an SMA/IPEX antenna interface, and we replaced the ceramic antenna with a patch antenna to reduce the weight of the system.

In order to realize the communication function independently from the smart phone, we chose the WIFI module instead of the Bluetooth module to ensure that the user can connect the signal only by carrying the IBGS. The WIFI module uses ESP8266, the volume of the module is only 11 mm × 10 mm, and the power supply current only needs 15 mA. It supports LWIP protocol, three communication modes: AP /STA /AP + STA coexistence mode, supports rich Smart AT commands, and can communicate with GD32 via UART serial port.

TABLE 2. Power consumption of the system.

Working parts	Rated working current (mA)		Average power consumption per second in 1 hours (mA)
	Run mode	Standby mode	
GD32	32.40	0.61	16.93
SNR9813	50	Continue working	50
Speaker	0.25		0.12
Microphone	5×10^{-5}		3.67×10^{-5}
Maix Nano	300		110
OV7740	60		22
VL53L0 V2	6	5×10^{-6}	2.20
ATGM336H	20		14.67
ESP8266	15		11
Total			226.92

In addition to assisting outdoor walking, we use NFC technology to develop a gap in existing research — payment function. We choose PN532 module, which has three working modes: command mode, automatic card number reading, and automatic block data reading. It supports compatible label cards such as Fudan F08, Mifare 550, and Mifare570, and can communicate with GD32 via UART serial port. In IBGS, PN532 works in Card emulation mode to achieve power supply through the RF domain of the contactless card reader, and does not require power supply from the host device.

The IGBT system is powered by 3.7 V, 400 mAh lithium battery. The size of the battery is 36.88 mm × 19.81 mm, and the thickness is only 5.63 mm, which meets the needs of the system for small volume. The IGBT is embedded with a TP4056 constant current charging chip, which is convenient for users to charge the glasses clip. The module has a mini USB interface and has an overcurrent protection function to ensure safe charging. On average, the IBGS will be woken up once every 82 seconds. The power consumption of the system is calculated as shown in Table 2. According to theoretical estimates, a fully charged glasses clip can continue to work for 1.76 hours.

C. SOFTWARE WORKFLOW

The software flow chart is shown in Fig. 6. IBGS can realize one-key boot function. The lithium battery stably outputs 3.3 V DC voltage to power the system through the voltage regulator chip. When the GD32 chip works normally, the battery power is detected first, and the monitoring unit and communication unit are initialized.

After receiving the user's speech command, the speech unit returns the recognition result to the MCU. In the program,

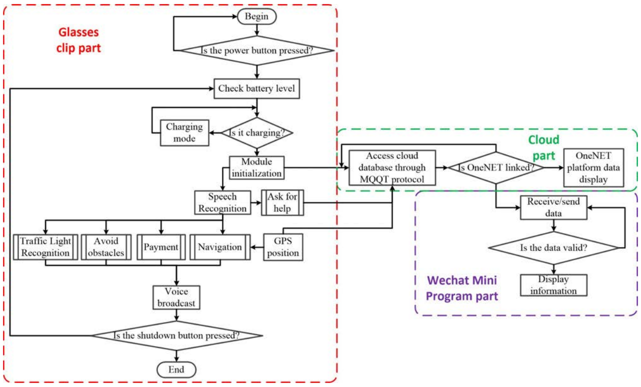


FIGURE 6. IBGS software flow chart. The speech recognition unit is the bridge between the MCU and the received data, it is also the transportation hub between the glasses clip and the outside world. The figure shows the data flow when the glasses clip works. The analyzed and processed data will be broadcast to the user or transmitted to the WeChat Mini Program through the cloud database.

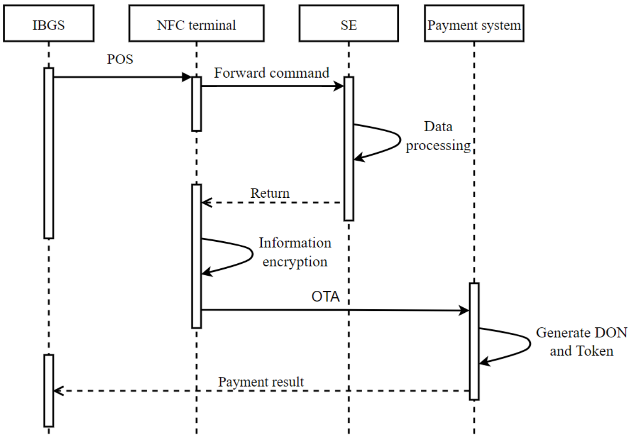
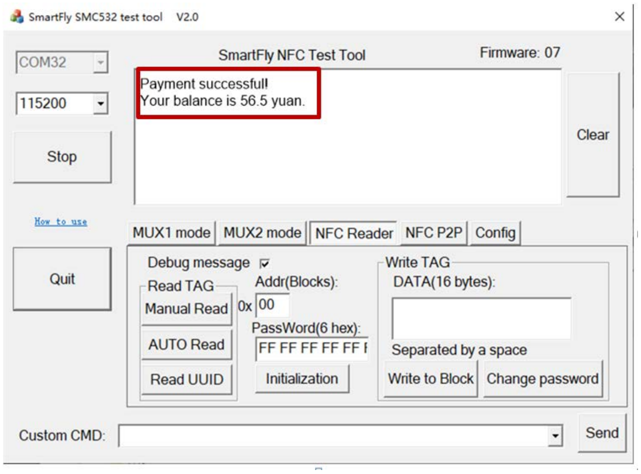


FIGURE 7. Sequence diagram for realizing NFC payment function. In mobile payment, NFC is used as a read device to complete the card simulation function, and it does not need power supply whereas generates a radio frequency field by itself.

we set flags for different functions. If it is identified as a traffic light function or obstacle avoidance function, change the flag of the image processing unit or the ranging unit, and then turn on their corresponding interrupts to receive data. The MCU analyzes the data, corresponds to the respective identification results, and corresponds them to the pre-set broadcast entries. For example, if the image processing unit returns data that represent a ''green light'' to the MCU, the MCU controls the speech unit to announce that the traffic light is green. Similarly, when the distance of the obstacle ahead received by the MCU is less than the set value, it will immediately control the speech broadcasting unit to issue a warning.

If the user wants to try the payment function, the speech command is used to open the NFC, then the NFC forwards the command from the external reader to the Secure Element (SE), which replies to the processed data through the NFC controller. After this information is encrypted, it will be sent to the payment system by the Over-the-Air (OTA) technology to generate the encrypted device account (DAN) and payment key (Token) of the transaction [37]. The transaction process is shown Fig. 7. In order to visualize the function, we use the host computer to simulate the payment process, as shown in Fig. 8.

If the navigation function is selected, the longitude and latitude information collected by GPS will be uploaded through WIFI, and Baidu map API will be called. MCU will convert it into Baidu coordinate system and receive feedback information on path information planning. Finally, MCU will send the information to the speech unit for broadcasting. At the same time, the movement track collected by GPS is uploaded to the OneNET cloud database through WIFI in real time. Before connecting to the database, the system connection status will be detected first. If it fails to connect to the network, the GD32 will first use the internal flash to store data, and the maximum data storage capacity can reach 1024 k. This process is shown in Fig. 9. WeChat Mini Program communicates with OneNET through the MQTT protocol, downloads longitude


SmartFly SMC532 test tool V2.0

SmartFly NFC Test Tool

Firmware: 07

COM32

115200

Payment successful!
Your balance is 56.5 yuan.

Debug message ☑

Addr(Blocks): 0x 00

PassWord(6 hex): FF FF FF FF FF F

DATA(16 bytes):

Separated by a space

Custom CMD:

FIGURE 8. Use the host computer to simulate the transaction process. During the actual payment, the information in the picture will be fed back to the MCU and then broadcast to the user.

and latitude information, and then calls Tencent map API to visualize longitude and latitude information.

WeChat Mini Program is an application that can be used without download and installation [38]. The WeChat Mini Program page of IBGS is shown in Fig. 10, which consists of the home page, the location page, and the help page. The home page is used for guardians to log in and connect to the cloud server. The location page displays the user's movement track. The help page displays the information sent by the user at the eyeglasses clip. After the guardian presses the ''Help'' button, it will automatically jump to the location page to display the user's location.

IV. DATA PROCESSING

A. RECOGNITION OF TRAFFIC LIGHTS

YOLOv3 [39] is a multi-target detection algorithm, which has been widely used in target detection due to its fast recognition

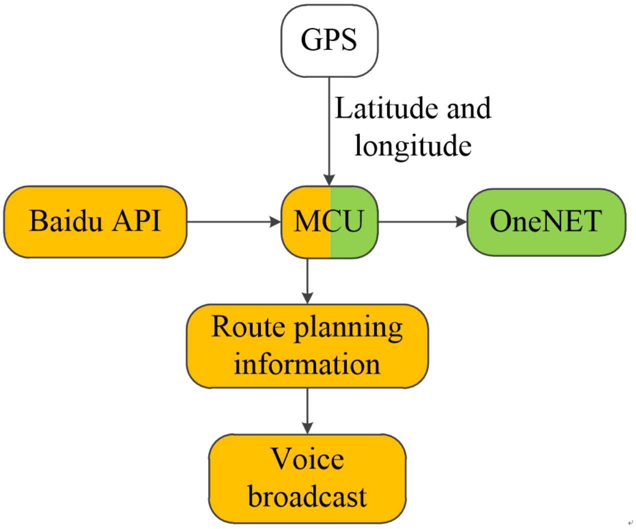
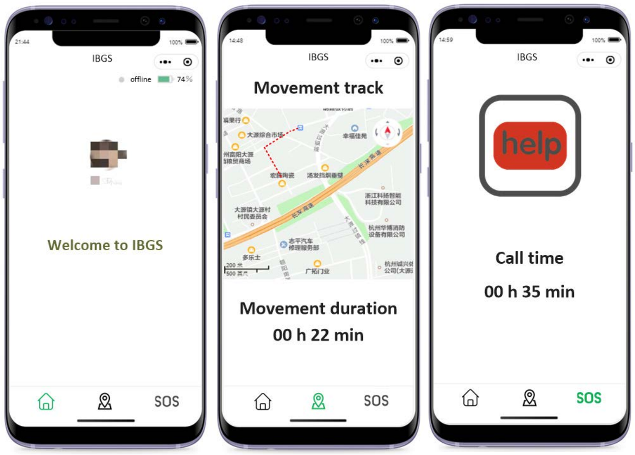


FIGURE 9. Flow of information collected by GPS.



(a) Home page (b) Location page (c) Help page

FIGURE 10. WeChat Mini Program page design.

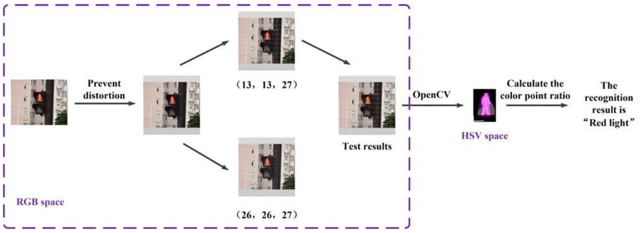


FIGURE 11. Flow of the YOLOv3-tiny algorithm. In this study, the dataset contains 4 categories, so the shapes of the prediction results are (13, 13, 27) and (26, 26, 27) respectively. In this example, the traffic lights to be identified are small, so 26 × 26 feature map is used for detection.

TABLE 3. Confusion matrix.

True label	Model prediction	
	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

speed and high accuracy [40]. Different from YOLOv3 which uses Darknet-53 as the backbone network, YOLOv3-tiny is used as the backbone network in this paper to pursue the recognition performance of lightweight and high speed, and improve the prediction speed whereas maintaining the prediction accuracy, which is suitable for building the model of checking traffic lights in daily life.

After the collected images are fed into the model, YOLOv3 controls the size of the output feature image by adjusting the convolution step. First, the image was converted into a 416 × 416 grid, and gray bars were added to prevent distortion. Then the multi-feature layer is extracted for target detection. The first feature graph was downsampled 32 times with a size of 13 × 13, and the second feature graph was downsampled 16 times with a size of 26 × 26. The output dimension is $(N, N, 3 \times (4 + 1 + N_{type}))$, where $N \times N$ represents the number of grids, and each grid has 3 anchors to assist with goal positioning. Each anchor has 1-dimensional prediction frame confidence and 4-dimensional prediction frame position information: x_offset (the offset of the upper-left corner of the grid relative to the X-axis), y_offset (the offset of the upper-left corner of the grid relative to the Y-axis), height, and width. N_{type} indicates the number of target object categories. According to the Feature Pyramid Network (FPN) architecture [41], the small size feature map is used to detect large-size objects, and the large size feature map is used to detect small size objects. Taking a traffic light image in training set as an example, the network structure is shown in Fig. 11.

When the model recognizes the target image, the location information on the image is returned, and OpenCV is used to extract the area, of which color is converted from RGB space to HSV space, and the proportion of each color pixel in this area is determined, so as to completely judge the status of the traffic signal light [42]. For IBGS functional requirements, we specially build a dataset of traffic lights and use Labellmg to mark them. The dataset includes 5,369 images of 4 kinds of traffic lights common to pedestrians in daily life. We use the 8 megapixels SONY IMX179 sensor to collect the pedestrian traffic indicators in different places in the city at different times and under different weather conditions to improve the recognition accuracy of the model. Of the more than 5,000 traffic light images, 3,221 of them are used as a training set, 1,074 as a validation set, and the rest as a test set. We use Intel (R) Core (TM) I5-8400 CPU with 3.20 GHz, CUDA11.3, and CUDNN7.6.5 for image processing training and detection under the win10_X64 system. A portion of the dataset is shown in Fig. 12.

In this paper, mean average precision (mAP) [43] is used to evaluate the target detection effect of the model. In order to calculate the mAP value, the Confusion Matrix of this model is given in Table 3.

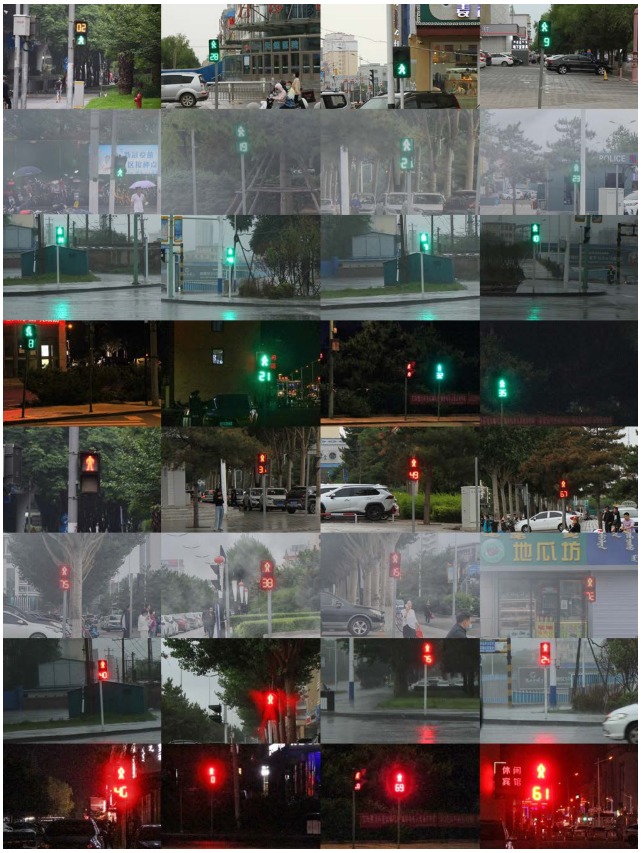


FIGURE 12. **Collected dataset under different weather conditions. The dataset includes photos of green and red lights on sunny, foggy, rainy, dark days.**

True Positive (TP) indicates that traffic lights are recognized and the classification of traffic lights is consistent with the actual situation. False Positive (FP) indicates that traffic lights are recognized but the classification of traffic lights is inconsistent with the actual situation. False Negative (FN) indicates that there are traffic lights in the image, but not correctly identified. The precision rate and recall rate are defined in (1) and (2) to evaluate the performance of the network.

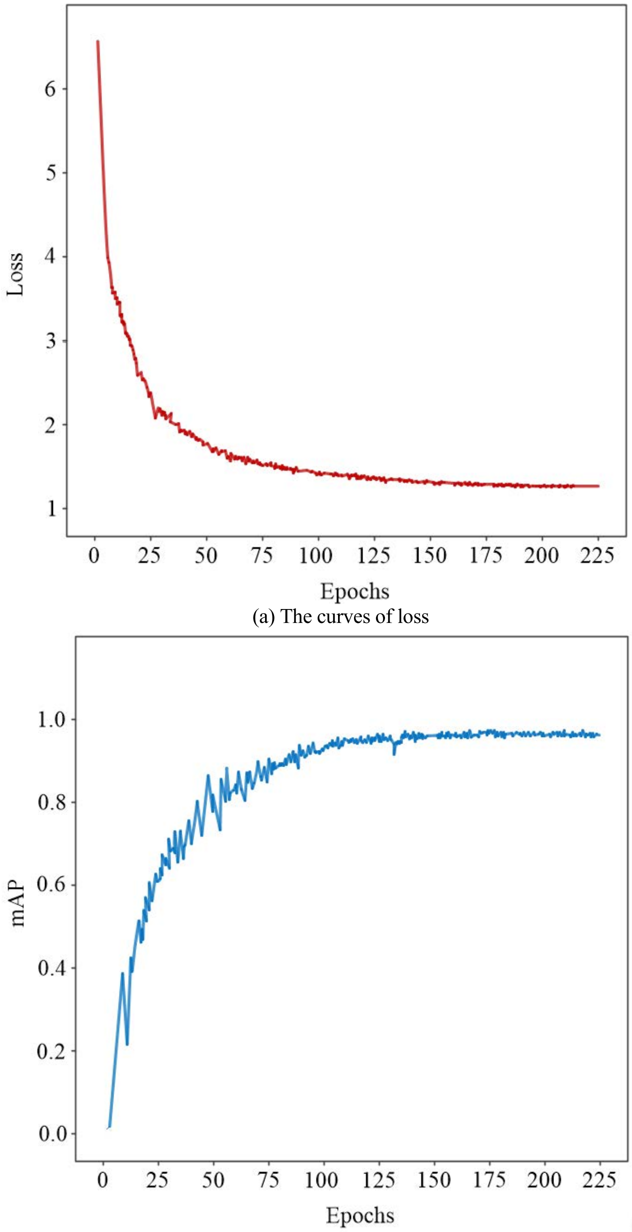
$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

Precision represents the accuracy of the prediction results, recall represents the ability to distinguish positive samples. The precision-recall curve can be obtained by taking recall as the abscissa and precision as the ordinate. The mAP value can be represented by the precision-recall curve and the area enclosed by the coordinate axes. The closer the mAP value is to 1, the higher the detection precision is.

Average loss and mAP after training are shown in Fig. 13.

It can be seen that within the first 30 epochs, loss rapidly converges. After 200 epochs, the loss has tended to be stable at 1.2931. At this point, the mAP of the model is roughly stable. After 225 epochs, the mAP was the highest, accounting for 97%. The corresponding training weight files and models were selected for traffic light recognition.



(a) The curves of loss

(b) The curves of mAP

FIGURE 13. **Curves of loss and mAP during the training process.**

B. THE REALIZATION OF HUMAN–COMPUTER SPEECH INTERACTION

The functionality of IBGS relies on speech interaction. Considering the high requirements of the system for stream speech recognition, ConvT-T is adopted in this paper, which has a lower delay, lower frame rate, and fewer parameters. ConvT-T combines RNN-T and Transformer, and a unidirectional Transformer with interleaved convolution [44] is used in audio encoder to reduce frame rate and obtain future information. In addition, in a continuous speech signal, there are

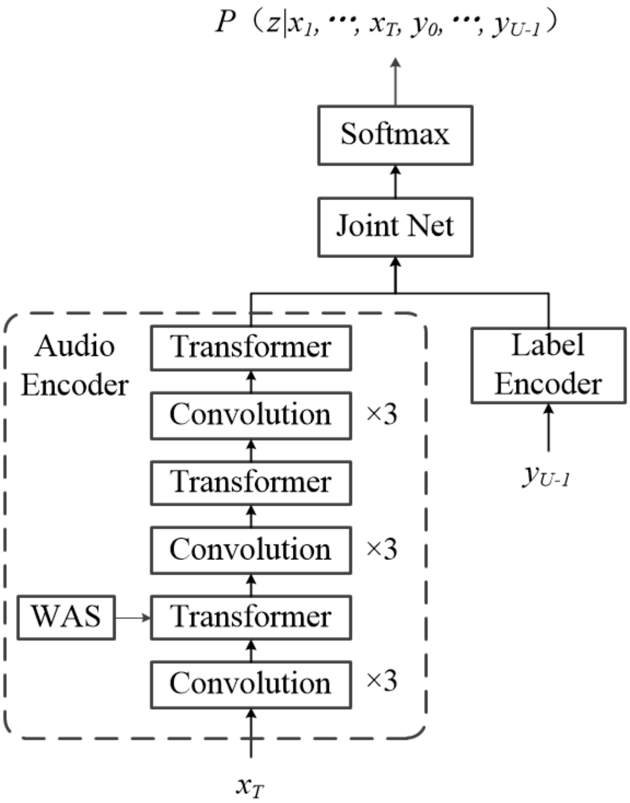


FIGURE 14. ConvT-T structure diagram that integrates WAS.

usually some elements that are not helpful for the recognition results, such as silence, or continuous acoustic units that share similarities, which are redundant for long-distance dependencies. In order to improve the efficiency of multi-head attention in the Transformer, we add WAS to the attention mechanism.

The following introduces the ConvT-T architecture. As shown in Fig. 14. The Transducer structure [28] uses a feed-forward network (FFN) and softmax to combine the output of the audio encoder and the output of the label encoder, thereby obtaining a probability distribution of the output label at each time node. Define the input sound sequence of T frame as $x = (x_1, \ldots, x_T)$, the transcription mark sequence as $y = (y_1, \ldots, y_U)$, and the input sound sequence is encoded by the encoder to generate the encoding state sequence. Thus, for each alignment path, the conditional distribution is defined as $P(z \mid x)$, and z represents the alignment path of the encoded features, which consists of blank labels and the actual output label sequence y. When the input sequence is known, the probability distribution of the output label is obtained by integrating the alignment relationship between the output label and the encoded feature, as shown in (3).

$$P(y|x) = \sum_{z \in Z(y,T)} P(z|x) \tag{3}$$

where $Z(y,T)$ is the set of effectively aligned label sequences of length T.

In Audio encoder, there are three sets of convolution layers combined with unidirectional Transformer. Input for self-attention in Transformer is obtained from the staggered convolution layer. Each combination of convolution layer and Transformer gradually downsample input audio sequence, so as to reduce the complexity of the training process whereas ensuring accuracy.

When the embedding vector is added to self-attention, it needs to be converted into $\boldsymbol{q}$, $\boldsymbol{k}$, $\boldsymbol{v}$, which represent a "query", a "key", and a "value" projection of each word in the input sentence respectively. The convolution layer multiples the embedding vector by W^Q, W^K, W^v through matrix operation and converts it into $\boldsymbol{q}$, $\boldsymbol{k}$, and $\boldsymbol{v}$ vectors, so that the information exchange between the input words occurs. The formula is shown as follows.

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{QK}^T}{d_k^{1/2}}\right)\boldsymbol{V} \tag{4}$$

where d_k is a dimension of the vector $\boldsymbol{q}$ or $\boldsymbol{k}$, $\boldsymbol{K}^T$ represents the transpose of $\boldsymbol{K}$.

A disadvantage of using the softmax function to get attention probabilities is that it often leads to dense attention. Therefore, we aim to induce sparsity in the attention probability distribution by dynamically determining a threshold for each time frame. All attention probabilities below this threshold are set to 0, and the remaining probabilities are renormalized. The threshold is essentially the difference between the mean and the standard deviation of the attention probability at a certain position in the query. It is worth noting that the user can specify a hyperparameter γ ($\gamma = 0.5$ in this paper) to multiply this standard deviation. The threshold is represented by θ_i, then (5) can be obtained.

$$\theta_i = \frac{1}{L} - \Upsilon\left(\frac{\sum_{j=1}^{L}\left(\alpha_{i,j} - \frac{1}{L}\right)^2}{L-1}\right)^{1/2} \tag{5}$$

where L is the length of the $\boldsymbol{k}$ in self-attention. The attention probability $\alpha_{i,j}$ is obtained by the softmax function.

WAS is added to the first layer Transformer to suppress attention to silence. Finally, the output of the audio encoder will be fed into a fully-connected feed-forward neural network with single hidden layer along with the prediction network. In the joint net, Rectified Linear Unit (ReLU) is adopted as the activation function.

In order to verify the performance of the proposed method, LibriSpeech ASR corpus was used for evaluation, and the features of 80-channel filter banks were calculated from a 25 ms window with a 10 ms step. To prevent overfitting and improve generalization during model training, feature augmentation [45] is employed and a Transformer learning rate schedule [46] is applied. The warm-up phase has a step size of 8 k, and the learning rate is scheduled to increase linearly from 0 to 2.5×10^{-4}. When held to 25 k, the exponential decays to 10^{-5}. To avoid overfitting, Additive Gaussian Noise ($\mu = 0$, $\sigma = 0.01$) is added to the time domain and velocity perturbations are applied [47].

In speech recognition, some words need to be replaced, deleted, or inserted in order to keep the recognized word

TABLE 4. WER comparison with previously model.

Model	#Params (M)	WER (%)	
		test-clean	test-other
FullAttn T-T [33]	139	2.4	5.6
LAS [45]	361	2.8	6.8
BLSTM [30]	79	3.1	7.4
AmTrf-24 [48]	81	3.1	7.1
AmTrf-24-WAS [36]	81	2.8	6.7
ConvT-T [34]	67	3.5	8.3
ConvT-T-WAS	73	3.2	7.9

sequence consistent with the standard word sequence. The total number of these words, divided by the percentage of the number of words in the standard word sequence, is the WER as shown in (6). (7) shows that the lower the WER is, the higher the accuracy is.

$$\text{Word Error Rate} = 100 \times \frac{\text{Insertions+Substitutions+Delections}}{\text{Total Words in Correct Transcript}} \tag{6}$$

$$\text{Accuracy} = 100 - \text{WER}\% \tag{7}$$

The proposed ConvT-T-WAS is compared with ConvT-T and some recently proposed models in Table 4.

As can be seen from Table 4, ConvT-T with WAS reduced WER by 0.4 with similar model sizes, and it achieves WER of 7.9% in test-other. Compared with the models commonly used at present, this model achieves competitive recognition accuracy.

V. EXPERIMENT

This section shows the IBGS prototype and the performance test of IBGS in actual use, including the traffic light recognition function test, speech recognition function test, and obstacle avoidance function test. All subjects were fully informed of the purpose of the experiment and signed a voluntary consent form. The protocol has been registered on the Experimental Committee of the University of Shanghai for Science and Technology, and the project number is 10-21-347-001, involving human experiments.

A. FUNCTIONAL TESTING IN DIFFERENT SCENARIOS

1) TRAFFIC LIGHT RECOGNITION FUNCTION TEST

Firstly, we verify the accuracy of IBGS's performance on the traffic lights collected in real time. We adopt 2 megapixels OV7740 as the sensor of IBGS. The volunteer is 170 cm tall and walks on the street for image collection. The crosswalk traffic light height is about 2-2.5 m. The test starts when the horizontal distance between the volunteer and the traffic light is 1.5-2 m. The validation dataset included more than 1,200 traffic light samples from pedestrian traffic lights commonly used in urban areas. The recognition rate is different under different environmental parameters, so these images evenly cover three time periods of morning, middle, and late, as well as three weather types of sunny, rainy, and foggy days, ensuring the representativeness and universality of the dataset.

TABLE 5. Test situation of traffic light recognition function.

Period	Weather	Number of samples	Missed detection / False detection	Accuracy (%)
Morning	Sunny	300	1/9	96.67
	Rainy	300	5/8	95.67
	Foggy	300	6/6	96
Middle	Sunny	300	2/7	97
	Rainy	300	2/8	96.67
	Foggy	300	4/7	96.33
Late	Sunny	300	8/6	95.33
	Rainy	300	15/8	92.33
	Foggy	300	14/5	93.67

Real-time detection through the camera and statistics according to the speech feedback results, the statistical results are shown in Table 5.

YOLOv3-tiny can well meet the requirements of fast recognition for embedded devices. When combined with OpenCV, IBGS can recognize traffic lights with 97% accuracy. It can be seen from the obtained data that the recognition rate under conventional images is higher. At night, due to the darker light, the black backplane is not obvious, which is not conducive to being separated from the environment, and the recognition rate is low, but it can be guaranteed to be above 92%. In extreme weather, the recognition rate is also reduced due to the influence of fog and light. In addition, we found that under the complex street scene, the recognition rate is slightly different from that under the empty scene, which is affected by the flow of people, buildings, and street signs. The recognition accuracy of this model in practical application is basically equal to that of mAP in training.

2) SPEECH RECOGNITION FUNCTION TEST

In practical wear applications, IBGS faces a more complex environment. The presence of background noise can make recognition performance different from when the model was trained. In order to test the realization effect of speech recognition function of different SNR, we selected three scenes of a park, shopping mall, and road for testing, corresponding to the environment of 30, 60, and above 60 dB respectively, and tested 300 times of each scene on average. Considering that different people's accents will also affect the recognition effect, in the speech recognition test, three volunteers from different hometowns were asked to test. IBGS uses a 52dB pickup, and the sound intensity of the instructions issued by the wearer are about 40-60dB. The effect of real-time speech wake-up and speech capture is judged through functional feedback. The volunteer testing process is shown in Fig. 15. In practical application, the sentence error rate (SER) is used to measure the acceptance of functional instructions, which is roughly the same as WER. If a word of a sentence is identified as an error, the sentence is considered to be identified as an error. Here, accuracy is 100 minus SER. The statistical results are shown in Table 6.

TABLE 6. Test situation of speech recognition function.

Testing scenarios	Test item	Testing frequency			Error times			SER (%)			Mean accuracy
		VOL 1	VOL 2	VOL 3	VOL 1	VOL 2	VOL 3	VOL 1	VOL 2	VOL 3	
Park(≤30dB)	Speech wake-up	100	100	100	5	9	9	5	9	9	92.33%
	Speech capture	100	100	100	6	8	10	6	8	10	92%
Shopping mall(≤60dB)	Speech wake-up	100	100	100	6	8	9	6	8	9	92.33%
	Speech capture	100	100	100	7	9	11	7	9	11	91%
Road(>60dB)	Speech wake-up	100	100	100	6	9	9	6	9	9	92%
	Speech capture	100	100	100	8	10	11	8	10	11	90.33%

VOL = volunteer

(a) In the park

(b) In the shopping mall

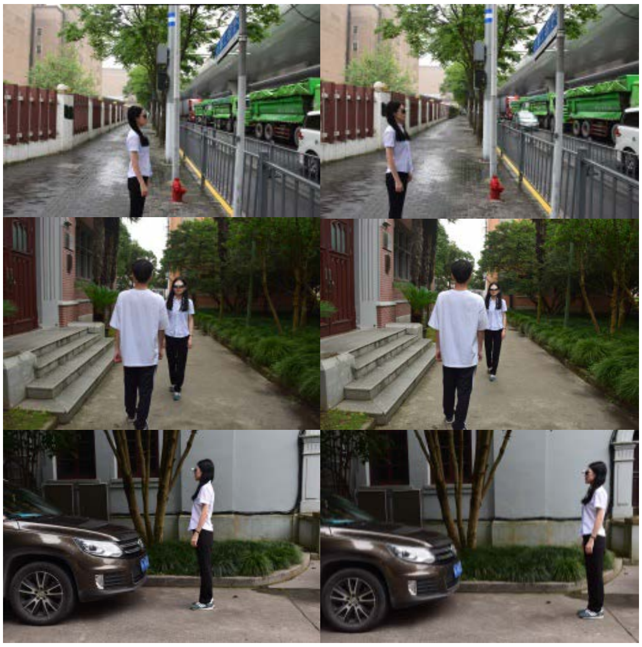
(c) On the road

FIGURE 15. Volunteers test speech recognition performance in different decibel scenes.

It can be seen that SER in practical application is slightly higher than WER in model training, because the judgment conditions are more stringent. Since white noise is added during model training, it has a certain anti-interference ability to noise in practical applications. In addition to white noise, the content of human voices in the background is also an important factor affecting the recognition effect. In the experiment, we found that the recognition effect in the area with more people will decrease. In addition, the human accent does have an impact on the recognition result, and it can be seen that the recognition result of volunteer 1 is slightly better than that of volunteer 3. However, no matter in which scene, the accuracy of speech capture can reach more than 90%. In addition, speech wake-up usually has no more than three marker words, so it performs better than speech capture, holding steady at around 92% accuracy.

3) OBSTACLE AVOIDANCE FUNCTION TEST

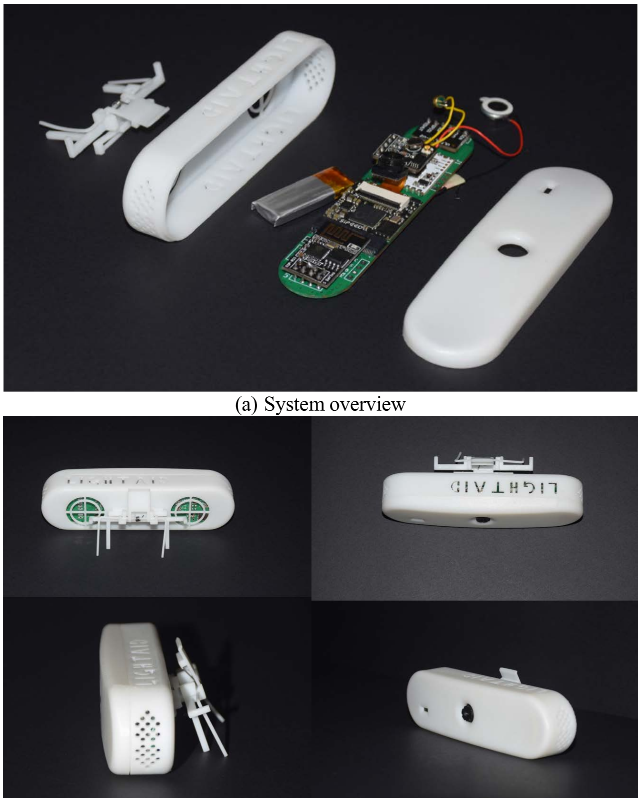
We tested three obstacles that blind people are most likely to encounter in daily outdoor walking: road signs, pedestrians, and vehicles. These three obstacles represent the three height levels of obstacles in daily outdoor walking. Because the VL53L0 V2 is a laser ranging sensor, the ambient light will interfere with the test results to a certain extent, so the test time is divided into two parts: day and night. The volunteer is 170 cm tall and walks on the street for testing. The volunteer testing process is shown in Fig. 16. IBGS will sound a speech alarm when the volunteer is 1 m away from the obstacle and judge the obstacle avoidance effect through real-time speech feedback. The statistical results are shown in Table 7.


(a) At safe distance

(b) At alarm distance

FIGURE 16. Volunteer conduct obstacle avoidance tests on various obstacles at different distances.

It can be seen that for road signs which are higher than the height of the volunteer, the detection accuracy of obstacles is 99%, and the detection accuracy of vehicles is 97.67%. For relatively small targets such as pedestrians, the detection accuracy can still reach more than 97%. For relatively low



(a) System overview

(b) Photo of glasses clip appearance

FIGURE 17. Prototype photos of the system.

TABLE 7. Test situation of obstacle avoidance function.

Test time	Test object	Testing frequency	Error times	Accuracy (%)
daytime	Street sign	300	1	99.67
	Pedestrian	300	8	97.33
	Vehicle	300	7	97.67
night	Street sign	300	2	99.33
	Pedestrian	300	10	96.67
	Vehicle	300	10	96.67

recognition targets such as vehicles, the VL53L0 V2 can still detect distances from multiple angles using distributed laser ranging. In addition, the recognition accuracy at night is slightly lower than that in the daytime, which may be because there are more interference factors such as neon lights at night, which affect the test level.

B. PROTOTYPE DISPLAY

Fig. 17 shows a photograph of a prototype of the proposed IBGS. The proposed system includes a spring clip, a shell, and a circuit board that integrates the sensors.

The overall volume of the glasses clip is 10 cm × 3 cm × 2.5 cm, and the weight is only 65 g, of which the shell is 35 g, and the spring clip and other components weigh 25 g, which is very convenient to carry. As shown in Table 8, the realization cost of the glasses clip is $40.3. As a multifunctional glasses clip, IBGS has an advantage in the field of low-cost wearable blindness assistance systems.

TABLE 8. Cost of the glasses clip.

Components	Price per unit ($)
3D printing	8.52
GD32F103VET6	3.08
Maix Nano	9.57
OV7740	2.24
SNR9813	1.78
Speaker	0.34
Microphone	0.30
VL53L0 V2	1.57
ATGM336H	5.16
PN532	2.54
ESP8266	1.91
TP4056	0.48
3.7V battery	2.06
Other	0.75
Total	40.30

After scene verification, the proposed glasses clip can well realize the function of helping the blind to go out independently, providing a new solution to assisting visually challenged.

VI. CONCLUSION AND FUTURE RESEARCH

This paper introduces an intelligent system which can help the blind to go out independently. The designed system has the characteristics of being low cost, lightweight, easy to carry, and can work independently from smartphones. It can realize the functions of speech interaction, traffic light recognition, laser obstacle avoidance, NFC payment, GPS navigation, and so on. Through IoT technology, the guardian can easily obtain the walking status of the blind at any time.

We demonstrate the feasibility of using the YOLOv3-tiny network and OpenCV for traffic light recognition of embedded mobile devices through experiments. The proposed method of ConvT-T with WAS reaches an accuracy of 92.4%. Compared with the currently popular models, the method has a competitive precision with fewer parameters.

In future work, we plan to share public benchmark datasets by extending the traffic light sample types. In addition, in terms of speech recognition algorithms, it is expected to reduce the number of parameters by optimizing the settings of the convolution layer, which means that the speed of speech recognition can be further improved. At the same time, we will also focus on solving the problem that there is no guarantee of WIFI signal all the time, by adding GSM or other modules to find an alternative solution when the WIFI signal is vacant.

ACKNOWLEDGMENT

(Kun Xia and Xueyong Li are co-first authors.)

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[2] G. Bradski, "The OpenCV library," *Softw. Tools Prof. Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[3] K. Finkenzeller, *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards Radio Frequency Identification and N Ear-Field Communication*. Hoboken, NJ, USA: Wiley, 2010.

[4] U. Masud, T. Saeed, H. M. Malaikah, F. U. Islam, and G. Abbas, "Smart assistive system for visually impaired people obstruction avoidance through object detection and classification," *IEEE Access*, vol. 10, pp. 13428–13441, 2022, doi: 10.1109/ACCESS.2022.3146320.

[5] D. Dharan, S. Kumar, and D. M. Getsy, "Visually impaired smart assistance," in *Proc. Int. Conf. Syst., Comput., Automat. Netw. (ICSCAN)*, 2021, pp. 1–3, doi: 10.1109/ICSCAN53069.2021.9526495.

[6] V. Nandalal, V. A. Kumar, A. Sujitha, G. Sumitha, and A. S. Sureka, "Intelligent multi-utility shoe for visually impaired persons," in *Proc. 2nd Int. Conf. Smart Electron. Commun. (ICOSEC)*, Oct. 2021, pp. 1102–1108, doi: 10.1109/ICOSEC51865.2021.9591728.

[7] V. Singh, S. Sindhu, and R. Arora, "BUZZFEET: Blind man shoes," in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput. (COMITCon)*, Feb. 2019, pp. 356–359, doi: 10.1109/COMITCon.2019.8862205.

[8] H. Huang, S. Huang, and J. Liu, "Intelligent blind travel navigation helmet," *China New Commun.*, vol. 22, no. 1, p. 51, 2020.

[9] F. Ashiq, M. Asif, M. B. Ahmad, S. Zafar, K. Masood, T. Mahmood, M. T. Mahmood, and I. H. Lee, "CNN-based object recognition and tracking system to assist visually impaired people," *IEEE Access*, vol. 10, pp. 14819–14834, 2022, doi: 10.1109/ACCESS.2022.3148036.

[10] N. Kumar and A. Jain, "Smart navigation detection using deep-learning for visually impaired person," in *Proc. IEEE 2nd Int. Conf. Electr. Power Energy Syst. (ICEPES)*, Dec. 2021, pp. 1–5, doi: 10.1109/ICEPES52894.2021.9699479.

[11] W.-J. Chang, L.-B. Chen, C.-Y. Sie, and C.-H. Yang, "An artificial intelligence edge computing-based assistive system for visually impaired pedestrian safety at zebra crossings," *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 3–11, Feb. 2021, doi: 10.1109/TCE.2020.3037065.

[12] A. S. Romadhon and A. K. Husein, "Smart stick for the blind using arduino, altrasonic sensor and android," *J. Phys., Conf.*, vol. 1569, no. 3, 2020, Art. no. 032088.

[13] S. Chandna and A. Singhal, "Towards outdoor navigation system for visually impaired people using YOLOv5," in *Proc. 12th Int. Conf. Cloud Comput., Data Sci. Eng.*, Jan. 2022, pp. 617–622, doi: 10.1109/Confluence52989.2022.9734204.

[14] A. Bhattacharya and V. K. Asari, "Wearable walking aid system to assist visually impaired persons to navigate sidewalks," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Oct. 2021, pp. 1–7, doi: 10.1109/AIPR52630.2021.9762132.

[15] P. Kramomthong, C. Pintavirooj, and M. P. Paing, "Smart cane for assisting visually impaired people and the blind," in *Proc. 13th Biomed. Eng. Int. Conf. (BMEiCON)*, Nov. 2021, pp. 1–5, doi: 10.1109/BMEiCON53485.2021.9745212.

[16] S. Mohapatra, S. Rout, V. Tripathi, T. Saxena, and Y. Karuna, "Smart walking stick for blind integrated with SOS navigation system," in *Proc. 2nd Int. Conf. Trends Electron. Informat. (ICOEI)*, May 2018, pp. 441–447, doi: 10.1109/ICOEI.2018.8553935.

[17] S. Chakraborty, A. Bhadra, S. Mondal, S. Prasad, and S. Chakraborty, "An intelligent and smart belt for the blind people," in *Advances in Smart Communication Technology and Information Processing* (Lecture Notes in Networks and Systems), vol. 165. Singapore: Springer, 2021.

[18] S. Martinez-Cruz, L. A. Morales-Hernandez, G. I. Perez-Soto, J. P. Benitez-Rangel, and K. A. Camarillo-Gomez, "An outdoor navigation assistance system for visually impaired people in public transportation," *IEEE Access*, vol. 9, pp. 130767–130777, 2021, doi: 10.1109/ACCESS.2021.3111544.

[19] S. U. Rao, S. Ranganath, T. S. Ashwin, and G. R. M. Reddy, "A Google glass based real-time scene analysis for the visually impaired," *IEEE Access*, vol. 9, pp. 166351–166369, 2021, doi: 10.1109/ACCESS.2021.3135024.

[20] M. Biswas, S. Chaki, F. Ahammed, A. Anis, J. Ferdous, A. M. Siddika, D. A. Shila, and L. Gaur, "Prototype development of an assistive smart-stick for the visually challenged persons," in *Proc. 2nd Int. Conf. Innov. Pract. Technol. Manage. (ICIPTM)*, Feb. 2022, pp. 477–482, doi: 10.1109/ICIPTM54933.2022.9754183.

[21] M. Ashrafuzzaman, S. Saha, N. Uddin, P. K. Saha, S. Hossen, and K. Nur, "Design and development of a low-cost smart stick for visually impaired people," in *Proc. Int. Conf. Sci. Contemp. Technol. (ICSCT)*, Aug. 2021, pp. 1–6, doi: 10.1109/ICSCT53883.2021.9642500.

[22] M. M. K. Peyal, Q. M. A. U. Haque, T. Tahiat, S. Habib, A. Noor, and A. A. M. Azad, "Inexpensive voice assisted smart eyewear for visually impaired persons in context of Bangladesh," in *Proc. IEEE Global Humanitarian Technol. Conf. (GHTC)*, Oct. 2021, pp. 43–50, doi: 10.1109/GHTC53159.2021.9612468.

[23] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[24] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2011.

[25] N. Moritz, T. Hori, and J. L. Roux, "Streaming end-to-end speech recognition with joint CTC-attention based models," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 936–943.

[26] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4774–4778.

[27] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S.-Y. Chang, K. Rao, and A. Gruenstein, "Streaming end-to-end speech recognition for mobile devices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6381–6385.

[28] A. Graves, "Sequence transduction with recurrent neural networks," *Comput. Sci.*, vol. 58, no. 3, pp. 235–242, 2012.

[29] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888.

[30] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang, C. Fuegen, G. Zweig, and M. L. Seltzer, "Transformer-based acoustic modeling for hybrid speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6874–6878.

[31] F. Zhang, Y. Wang, X. Zhang, C. Liu, Y. Saraf, and G. Zweig, "Faster simpler and more accurate hybrid ASR systems using wordpieces," in *Proc. Interspeech*, 2020, pp. 976–980.

[32] C. F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. L. Seltzer, "Transformer-transducer: End-to-end speech recognition with self-attention," 2019, *arXiv:1910.12977*.

[33] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 7829–7833.

[34] W. Huang, W. Hu, Y. T. Yeung, and X. Chen, "Conv-transformer transducer: Low latency, low frame rate, streamable end-to-end speech recognition," 2020, *arXiv:2008.05750*.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[36] Y. Shi, Y. Wang, C. Wu, C. Fuegen, F. Zhang, D. Le, C. F. Yeh, and M. L. Seltzer, "Weak-attention suppression for transformer based speech recognition," 2020, *arXiv:2005.09137*.

[37] V. Njebiu, M. Kimwele, and R. Rimiru, "Secure contactless mobile payment system," in *Proc. IEEE Latin-Amer. Conf. Commun. (LATINCOM)*, Nov. 2021, pp. 1–6, doi: 10.1109/LATINCOM53176.2021.9647831.

[38] L. Hongwei, "Application of WeChat mini program," *Wireless Internet Technol.*, vol. 23, pp. 11–12 and 40, Dec. 2016.

[39] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[40] B. Sui, P. Zhang, and X. Wang, "An improved YOLOv3 gesture recognition algorithm," *J. Hebei Univ. Sci. Technol.*, vol. 42, no. 1, pp. 22–29, 2021.

[41] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[42] C. Yu and Q. Zhu, "Research on traffic light recognition algorithm based on deep learning and OpenCV," *Shanghai Auto*, vol. 7, pp. 19–22, Jul. 2019.

[43] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2006, pp. 11–18.

[44] T. Zhang, G. J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions for deep neural networks," 2017, *arXiv:1707.02725*.

[45] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[47] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[48] C. Wu, Y. Wang, Y. Shi, C.-F. Yeh, and F. Zhang, "Streaming transformer-based acoustic models using self-attention with augmented memory," in *Proc. Interspeech*, 2020, pp. 2132–2136.

KUN XIA received the B.Eng. degree in industrial automation and the Ph.D. degree in power electronics and power drives from the Hefei University of Technology (HFUT), Hefei, China, in 2002 and 2007, respectively. From 2007 to 2011, he was a Lecturer with the University of Shanghai for Science and Technology (USST), Shanghai, China. From 2011 to 2019, he was an Associate Professor and a Department Head with the Electrical Engineering Department, USST. From 2015 to 2016, he was also a Visiting Scholar with the Electrical and Computer Engineering Department, National University of Singapore, Singapore. Since 2020, he has been a Professor and a Vice President with the College of Innovation and Entrepreneurship, USST. He has also been an in charge of more than 50 research projects from the government and companies. He has published more than 80 articles. His research interests include motor and motor control and new energy application. He won the third prize of Scientific and Technological Progress Award of Zhejiang and Shanghai, in 2017 and 2020, respectively.

XUEYONG LI was born in China, in 1998. She received the B.Eng. degree from the Department of Electrical Engineering, University of Dalian Jiaotong University, Liaoning, China, in 2020. She is currently pursuing the M.Eng. degree with the Department of Electrical Engineering, University of Shanghai for Science and Technology. Her current research interests include deep learning and embedded chips.

HAIYANG LIU was born in China, in 1996. He received the B.Eng. degree from the Department of Electrical Engineering, University of Shanghai for Science and Technology (USST), Shanghai, China, in 2020, where he is currently pursuing the M.Eng. degree. His current research interest includes wireless communication.

MINGLI ZHOU was born in Anhui, China, in 1997. He received the B.S. degree in electrical engineering and automation from the School of Control Engineering, Chengdu University of Information Technology, Chengdu, China, in 2020. He is currently pursuing the master's degree in power electronics and power drives with the Laboratory of Electrical Engineering, University of Shanghai for Science and Technology, Shanghai, China. His current research interests include NNPP converter and deep learning.

KEXIN ZHU was born in China, in 1998. She received the B.Eng. degree from the Department of Automation, China University of Petroleum (Beijing), Beijing, China, in 2020. She is currently pursuing the M.Eng. degree with the Electrical Engineering Department, University of Shanghai for Science and Technology. Her current research interests include deep learning and industrial control systems.