IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 7, 2020, accepted March 20, 2020, date of publication March 31, 2020, date of current version April 16, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2984718*

# Unifying Visual Localization and Scene Recognition for People With Visual Impairment

**RUIQI CHENG[1], KAIWEI WANG[1], JIAN BAI[1], AND ZHIJIE XU[2], (Member, IEEE)**
[1]National Engineering Research Center of Optical Instrumentation, Zhejiang University, Hangzhou 310027, China
[2]School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, U.K.

Corresponding author: Kaiwei Wang (wangkaiwei@zju.edu.cn)

This work was supported by the grant from ZJU-Sunny Photonics Innovation Center (No. 2020-03) and the National Engineering Research Center of Optical Instrumentation.

**ABSTRACT** With the development of computer vision and mobile computing, assistive navigation for people with visual impairment arouses the attention of research communities. As two key challenges of assistive navigation, ''Where am I?'' and ''What are the surroundings?'' are still to be resolved by taking advantage of visual information. In this paper, we leverage the prevailing compact network as the backbone to build a unified network featuring two branches that implement scene description and scene recognition separately. Based on the unified network, the proposed pipeline performs scene recognition and visual localization simultaneously in the scenario of assistive navigation. The visual localization pipeline involves image retrieval and sequence matching. In the experiments, different configurations of the proposed pipeline are tested on public datasets to search for the optimal parameters. Moreover, on the real-world datasets captured by the wearable assistive device, the proposed assistive navigation pipeline is proved to achieve satisfactory performance. On the challenging dataset, the top-5 precision of scene recognition is more than 80%, and the visual localization precision is over 60% under a recall of 60%. The related codes and datasets are open-source online at https://github.com/chengricky/ScenePlaceRecognition.

**INDEX TERMS** Visual place recognition, global image descriptor, scene classification, assistive navigation.

## I. INTRODUCTION

In the era of artificial intelligence, different kinds of intelligent systems are emerging into the boom, including service robots, autonomous vehicles, and security surveillance systems, etc. The inspiring improvements have also occurred in the field of assistive technology, and various wearable devices have been developed by both the industrial and the academia to promote the living level of visually impaired people [1], [2]. However, the majority of those devices are aimed at obstacle avoidance or pathway detection. Indeed, those demands are the most critical demands for visually impaired people traveling both outdoors and indoors, whereas there are more functionalities to be studied for assistive navigation. In this paper, we focus on the visual localization and scene perception problems, which have not been deeply researched among the community of assistive technology. Deep convolutional networks have become powerful approaches to solve different kinds of computer vision tasks. Based on various networks, scene recognition (classification), as a classical computer vision task, has been studied thoroughly [3], [4]. The deep convolutional networks can also be used as descriptor extractors on the task of visual localization [5]. Therefore, it is possible to unify scene recognition and scene description in a single convolutional network, which resolves the two challenges of assistive navigation *Where am I?* and *What are the surroundings?* simultaneously.

As shown in Figure 1, the proposed convolutional network unifies the functionalities of scene recognition and visual localization. Specifically, the network consists of a preceding backbone defined as BaseNet and two subsequent branches that are represented as a blue branch for scene recognition and an orange branch for scene description. The blue branch generates the scene class of the input image, and the scene recognition result provides with the basic information of the surroundings. The orange branch extracts the NetVLAD (network-based vector of locally aggregated descriptors) descriptor [5] from the input image, and the global descriptor is used to search the localization result from the database images. The unified network yields both the scene class of the input image, which can be conveyed to the user through interaction approaches, and the scene descriptor

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

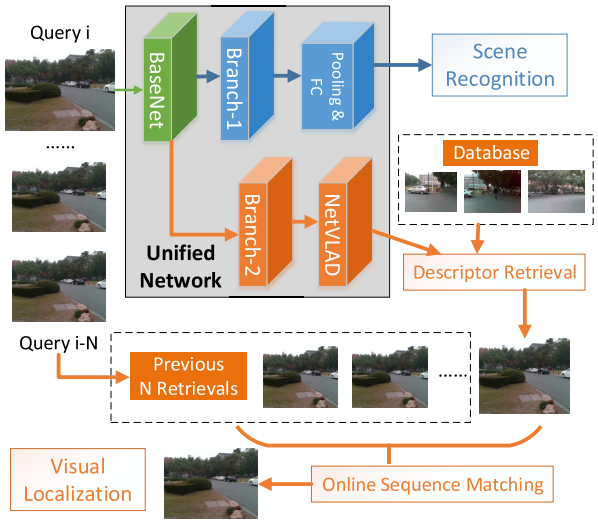


**FIGURE 1.** The proposed unified network and visual localization pipeline for assistive navigation. Branch-1 and Branch-2 share the same network structure but have different parameters.

of the input image. As shown in Figure 1, the descriptors of the query images are matched with the database descriptors by the fast retrieval of nearest neighbors. Subsequently, the proposed online sequence matching utilizes the retrieval results of consecutive frames to produce the best-matching database image, which is the result of visual localization.

The visual perception tasks in the scenarios of assistive navigation are more challenging than those in the fields of robotics or autonomous vehicles. The wearable camera is prone to capture images with motion blur and at low resolution, which causes difficulties for fine-grained scene understanding. Apart from that, multiple visual variations between query and database images (e.g. illumination, season, viewpoint, and dynamic objects) impede the robustness of visual localization. Taking those factors into consideration, we collect and label a real-world dataset in the scenario of assistive navigation to validate the proposed pipeline comprehensively.

The contributions of this paper lie in the two-fold aspects.

- The unified network is proposed to achieve both scene recognition and scene description simultaneously. The compressed deep descriptors are leveraged in the proposed visual localization pipeline composed of image retrieval and online sequence matching.
- Based on the real-world scenarios, the configurations of the unified network and the parameters of the visual localization pipeline are tuned. Both on the public and the self-collected datasets, the proposed pipeline is validated for assistive navigation.

The subsequent parts of this paper are arranged as follows. Section II reviews the state of the art on assistive navigation and the related convolutional network approaches on scene recognition and visual localization. Section III presents the details of the unified network for scene description and scene recognition and also presents the visual localization pipeline involving image retrieval and sequence matching. The network training procedures, the real-world datasets, and the corresponding assistive system and the experimental results are demonstrated in section IV. Section V concludes this paper.

## II. RELATED WORK

With the booming development of artificial intelligence and mobile Internet, the research community pays more and more attention to computer vision-based solutions on different issues [6], [7]. At the same time, the computer vision-based assistive navigation for visually impaired people is thriving. In this section, we focus on the work on scene recognition and visual localization, which are two of the most important applications in this field.

### A. ASSISTIVE SCENE RECOGNITION

There are plenty of contributions dedicated to object detection in the field of assistive navigation, such as generic object detection [8], staircase detection [9], and road barrier recognition [10]. In order to inform the visually impaired with ambient objects, Mekhalfi *et al.* [11] proposed a multi-label scene recognition algorithm based on compressive sensing, which works well at those places once visited. Yang *et al.* [12] seized pixel-wise semantic segmentation to cover navigation-related perception needs in a unified way and integrated the approach in a wearable navigation system by incorporating robust depth segmentation. Apart from the generic scene perception, we previously achieved assistive navigation at urban traffic intersections for visually impaired people [13]. The AECA (adaptive extraction and consistency analysis) algorithm detects the position and orientation of zebra crosswalks in real time [14]. The pedestrian crossing lights detection algorithm leverages candidate extraction, candidate recognition, and temporal-spatial analysis to implement robust performance in challenging scenarios [15].

In this paper, we concentrate on scene recognition instead of object detection or semantic segmentation, because scene recognition features lower computational complexity and is more suitable for interaction. There are different datasets used for training the scene recognition network, e.g. Places dataset [4] and SUN dataset [3]. Based on those datasets, a series of classical convolutional networks, such as AlexNet [16], VGGNet [17], GoogLeNet [18] and ResNet [19], have been proposed successively for scene recognition task. Moreover, different compact classification models, such as SqueezeNet [20], MobileNet [21] and ShuffleNet [22], are suitable for the mobile devices with limited resources. Based on those compact networks, a lot of applications, such as object detection and image segmentation [23], have been developed.

### B. (ASSISTIVE) VISUAL LOCALIZATION

The retrieval-based visual localization has attracted the attention of the research community of autonomous systems.

Retrieval-based methods obtain the best-matching database image of the query image and take that database image as the localization result. Localization methods based on image retrieval are suitable for assistive navigation in large-scale and dynamic environments, in that it does not require the precise metric maps.

To solve the cross-season visual changing of place recognition, Neubert *et al.* [24] proposed an appearance change prediction method based on the vocabularies of superpixels. Pepperell *et al.* [25] augmented the traditional one-dimension database with a directed graph, and used particle filter to achieve place recognition in networked environments. Abdollahyan *et al.* [26] presented a sequence-based approach to visual localization using the partial order kernel and the pre-trained CNN (convolutional neural network) descriptor. Maddern *et al.* [27] proposed illumination invariant transformation to improve visual localization performance during daylight hours. Those methods achieved superior performance on the season or illumination changes but did not pay attention to resolving more challenging visual variations of place recognition, such as dynamic objects or viewpoint changes. Ciarfuglia *et al.* [28] introduced a new bag-of-words loop closure detection by adopting a set of visual words weights learned offline accordingly to a discriminative criterion. Stumm *et al.* [29] presented a unified framework for defining, modelling and recognizing places in a way which is directly related to the underlying structure of features in the environment. Cascianelli *et al.* [30] utilized the landmark-based CNN descriptor and the covisibility graph to preserve scene geometric and semantic structure and to improve appearance invariance. Those methods achieved satisfactory visual localization performance on real-world datasets.

Currently, the research community of assistive technology tends to solve localization issues in indoor scenarios. Meanwhile, most of the approaches are based on beacon-based navigation systems [31], [32], where the hardware maintenance at the specific places, even if the maintenance cost can be controlled, limits the application range of navigation systems. In our previous work, we proposed a series of visual localization approaches in the field of assistive navigation. The key position prediction algorithm [33] uses conventional image descriptors based on multimodal images and GNSS (global navigation satellite system) data to localize the users at the user-defined key positions. Subsequently, we proposed the improved localization approach OpenMPR [34], where the off-the-shelf CNN descriptors along with other multimodal descriptors are optimized in the sequence matching pipeline by the genetic algorithm. Hu *et al.* [35] introduced the panoramic annular camera to visual odometry so as to robustify the positioning and mapping performance in the assistive navigation.

However, the previous work on assistive navigation only implemented one of scene perception and visual localization. The scene description and classification are related in terms of assistive functionality and network structure. Meanwhile, considering the limited resources of mobile devices, it is necessary to unify scene recognition and visual localization in a single network in order to shrink the computational complexity.

## III. METHODOLOGY

In this section, the proposed pipeline for assistive navigation and the details of the proposed network are illustrated. The proposed network outputs both scene recognition and scene description, which are leveraged in scene perception and visual localization respectively. Utilizing the scene descriptors, the assistive navigation pipeline yields visual localization results by fast retrieval and sequence matching.

### A. UNIFIED NETWORK

In this paper, the unified convolutional network is proposed to generate the scene class and the scene descriptor simultaneously. In order to make the whole system run smoothly on mobile devices with limited resources, we leverage the compact convolutional neural network to achieve our goal. In this paper, we consider two kinds of efficient neural networks: *MobileNet V2* [23] and *ShuffleNet V2* [36]. The backbone structure of MobileNet V2 / ShuffleNet V2 corresponds to BaseNet and Branch-1 (or Branch-2) in Figure 1. Herein, we present the characteristics and the structures of the two networks.

#### 1) BACKBONE NETWORK: MobileNet V2

Based on the preceding MobileNet that utilizes depth-wise separable convolutions to reduce the computational complexity, MobileNet V2 promotes the network performance and compresses the network size by (addition operation) shortcut structures and inverted residuals with linear bottlenecks. Different from standard convolutions, depth-wise separable convolutions involve depth-wise convolutions and point-wise convolutions, thus FLOPs (the number of floating-point multiplication-adds) shrink to around $1/K^2$ of the standard convolutions assuming the kernel size is $K \times K$.

In Table 1, we list the layer structure of a basic block (inverted residual with linear bottleneck) in MobileNet V2. The inverted residual means that the shortcut of adjacent blocks occurs at the feature maps with small depth, thus improving memory efficiency. That is to say, the intermediate feature maps in the block have a large depth width, which is controlled by the expansion factor $E$. In those blocks where $E = 1$, the layer 1 of Table 1 is omitted. The linear bottleneck gets rid of the activation layer of the last point-wise

**TABLE 1. The inverted residual with linear bottleneck in MobileNet V2.**

| Index | Layer | Output Size |
|---|---|---|
| 0 | Input | $h \times w \times d$ |
| 1 | Conv1_1 - BN - ReLU6 | $h \times w \times (E * d)$ |
| 2 | DWConv1_s - BN - ReLU6 | $(h/s) \times (w/s) \times (E * d)$ |
| 3 | Conv1_1 - BN | $(h/s) \times (w/s) \times d'$ |

ConvN_M denotes the convolutional layer with the kernel of N by N and the stride of M, DWConv denotes depth-wise convolution, BN denotes batch normalization layer, and E denotes expansion factor.

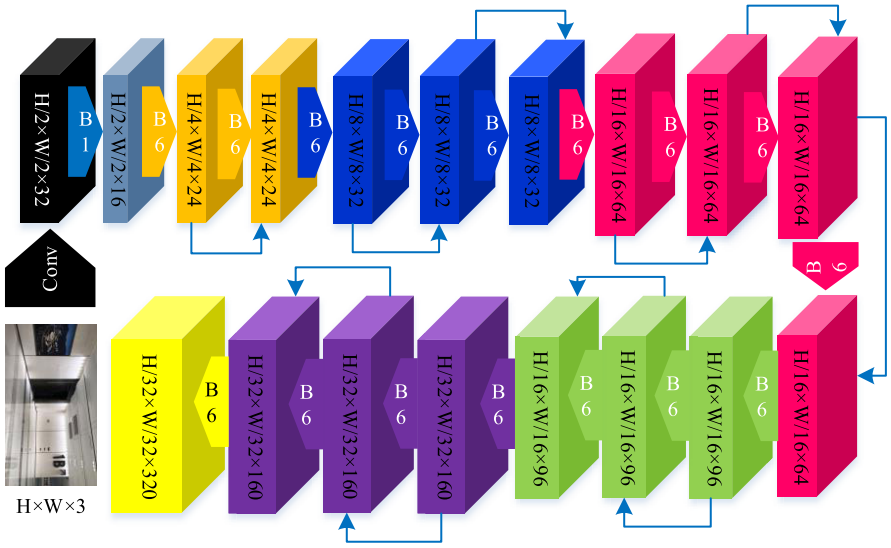


**FIGURE 2.** The backbone structure of MobileNet V2 used in the proposed unified network. The linear bottleneck with inverted residuals is denoted as "B n", where n is the channel expansion factor of intermediate layers in the bottleneck block. The shortcuts between the adjacent layers are presented with arrows.

convolution in the bottleneck, which prevents features from being disturbed by the nonlinear layer (e.g. ReLU).

As shown in Figure 2, the backbone structure of MobileNet V2 is composed of one convolutional layer (with a batch normalization layer and a ReLU6 activation layer) and successive 17 inverted residual blocks. Those blocks feature different expansion factors, different output depth sizes, and different convolution strides. Taking the image with the size of $H \times W \times 3$ as input, the backbone part of MobileNet V2 generates a feature map with the size of $H/32 \times W/32 \times 320$.

### 2) BACKBONE NETWORK: ShuffleNet V2

Another choice of the backbone is the prevailing efficient network ShuffleNet V2, which is optimized based on ShuffleNet according to four guidelines, including equal channel width for input and output layers, moderate group convolution, less network fragmentation, and limited element-wise operations. The basic block of ShuffleNet V2 is presented in Figure 3, which is similar to the inverted residual blocks of MobileNet V2, but the intermediate feature maps are not expanded in channel width. Moreover, the input feature maps are split in the channel dimension into two branches, which undergo different operations. Meanwhile, channel shuffle that rearranges the order of channels in feature maps makes the two branches of the basic block in ShuffleNet V2 communicate information. As concatenation exists in each basic block, there are more shortcuts in ShuffleNet V2 compared with MobileNet V2.

As shown in Table 2, the backbone structure of ShuffleNet V2 is composed of a convolutional layer (with a batch normalization layer and a ReLU activation layer), a max-pooling layer, and successive 16 basic blocks. Those blocks feature

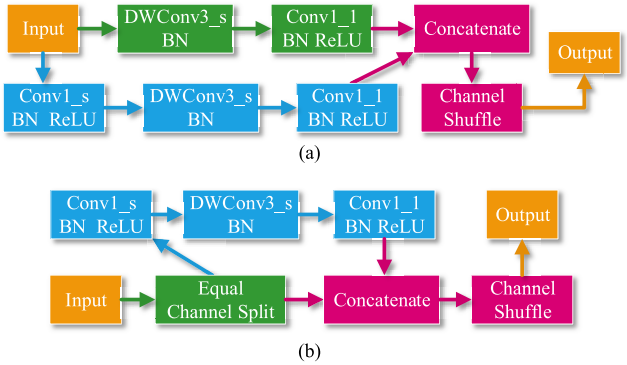


**FIGURE 3.** The basic block of ShuffleNet V2. (a) The basic block with the spatial down-sampling output whose depth is larger than that of the input. (b) The basic block in which the input and output have the same size.

different output depth sizes and different convolution strides. Taking the image with the size of $H \times W \times 3$ as input, the backbone part of ShuffleNet V2 generates a feature map with the size of $H/32 \times W/32 \times 464$.

### 3) BRANCH OF SCENE RECOGNITION

In the unified network, the scene recognition branch (shown as blue in Figure 1), together with the backbone structure, composes an intact scene recognition pipeline. Based on the backbone networks presented above, the scene recognition branch is as Table 3. As we use Places-365 dataset [4] as the training dataset of scene perception, the number of output classes is 365.

### 4) BRANCH OF SCENE DESCRIPTION

Based on the backbone network, the branch of visual localization is proposed to extract the scene descriptors of input

**TABLE 2. The backbone structure of ShuffleNet V2 (In this paper, D = 116).**

| Index | Layer | Output Size |
|---|---|---|
| 0 | Input | $H \times W \times 3$ |
| 1 | Conv | $H/2 \times W/2 \times 24$ |
| 2 | MaxPooling | $H/4 \times W/4 \times 24$ |
| 3 | Block1-1 | $H/8 \times W/8 \times D$ |
| 4 | Block1-2 | $H/8 \times W/8 \times D$ |
| 5 | Block1-3 | $H/8 \times W/8 \times D$ |
| 6 | Block1-4 | $H/8 \times W/8 \times D$ |
| 7 | Block2-1 | $H/16 \times W/16 \times 2D$ |
| 8 | Block2-2 | $H/16 \times W/16 \times 2D$ |
| 9 | Block2-3 | $H/16 \times W/16 \times 2D$ |
| 10 | Block2-4 | $H/16 \times W/16 \times 2D$ |
| 11 | Block2-5 | $H/16 \times W/16 \times 2D$ |
| 12 | Block2-6 | $H/16 \times W/16 \times 2D$ |
| 13 | Block2-7 | $H/16 \times W/16 \times 2D$ |
| 14 | Block2-8 | $H/16 \times W/16 \times 2D$ |
| 15 | Block3-1 | $H/32 \times W/23 \times 4D$ |
| 16 | Block3-2 | $H/32 \times W/32 \times 4D$ |
| 17 | Block3-3 | $H/32 \times W/32 \times 4D$ |
| 18 | Block3-4 | $H/32 \times W/32 \times 4D$ |

**TABLE 3. The pooling and FC layers of the scene recognition branch.**

| Backbone | Layer | Output Size |
|---|---|---|
| MobileNet V2 | Conv1_1 - BN - ReLU6 | $h \times w \times 1280$ |
| | AvePool | $1 \times 1 \times 1280$ |
| | Dropout - Linear | $1 \times 1 \times 365$ |
| ShuffleNet V2 | Conv1_1 - BN - ReLU | $h \times w \times 1024$ |
| | AvePool | $1 \times 1 \times 1024$ |
| | Linear | $1 \times 1 \times 365$ |

AvePool denotes average pooling layer.

images. NetVLAD is an efficient approach to pooling discriminative features from the preceding feature maps. Herein, we summarize it in brief. Obtained from the backbone structure (i.e. Branch-2 in Figure 1), the feature maps $F_0$ with the size of $W \times H \times D$ can be viewed as column-like local features, which are defined as

$$\{F_0^i \in \mathbb{R}^D | i = 1, 2, 3, \ldots, N; N = W \times H\}. \quad (1)$$

After column-wise L2-normalization, the features are denoted as $F_n$, and are assigned to the cluster centers $\{C_k \in \mathbb{R}^D | k = 1, 2, 3, \ldots, K\}$ proportional to their proximity. This operation is called soft assignment, the coefficient of soft assignment is represented as

$$a_k(F_n^i) = \frac{\exp(-\alpha||F_n^i - C_k||^2)}{\sum_{k'=1}^{K} \exp(-\alpha||F_n^i - C_{k'}||^2)} \quad (2)$$

$$= \frac{\exp(2\alpha C_k F_n^i - \alpha||C_k||^2)}{\sum_{k'=1}^{K} \exp(2\alpha C_{k'} F_n^i - \alpha||C_{k'}||^2)}. \quad (3)$$

In this paper, the cluster number $K$ is set as 64. Apparently, the soft assignment operation can be implemented by a convolutional layer with a softmax layer. Finally, the pooling results are defined as $\{V_k \in \mathbb{R}^D | k = 1, 2, 3, \ldots, K\}$, where

$$V_k = \sum_{i=1}^{N} a_k(F_n^i)(F_n^i - C_k). \quad (4)$$

Having L2-normalized each $D$-dimensional $V_k$ vector, we concatenate those vectors into a global descriptor, which is then L2-normalized globally.

### B. PIPELINE OF ASSISTIVE NAVIGATION

In this section, the pipeline of visual localization with scene recognition is demonstrated. For database images, the network inference is executed in prior, and the scene descriptors are saved. In the online phase, the query image is also fed into the network to yield the descriptor and the scene class. The scene recognition is conveyed to the users directly as the scene perception results, and the scene descriptors are used in the image retrieval of visual localization. The randomized k-d forest is utilized to retrieve the top-K candidates of the query image from database images.

In order to robustify the visual localization further, sequence matching determines the final result for each query image. The online sequence matching proposed by our previous work OpenMPR is modified to adapt to this work. In our previous work, the sequence matching only considers the nearest neighbor of the query. In this paper, the similarity matrix that is formed by the similarity value (the reciprocal of Euclidean distance) of top-K retrieved neighbors is leveraged for sequence matching. In Figure 4, each row of the matrix represents a query image, meanwhile, each column represents a database image. The similarity matrix is a sparse matrix, because only the retrieved top-K database images of a query image are assigned with similarity values and others are assigned with zero.

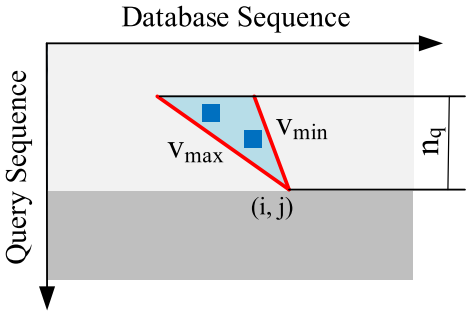


**FIGURE 4. The online sequence matching strategy used in this paper. The matching score of a database-query pair is determined by the top-K nearest neighbors' distance in its associated cone region.**

As shown in Figure 4, an online cone-based searching is carried out upon every query-database pair. Taking the query-database pair $(i, j)$ as the vertex, the corresponding cone-like searching region is limited by sequential length $n_q$, maximal velocity $v_{max}$ and minimal velocity $v_{min}$. Different from the offline searching algorithm in SeqSLAM [37], the online one only makes use of the past query images, regardless of the future query images. Within the cone region, the sum of similarity values is defined as $sum_{match}$. Then, the score $s_{i,j}$ of any query-database pair $(i, j)$ is defined as

$$s_{i,j} = \frac{sum_{match}}{n_q}. \quad (5)$$

For each query image, the latent best-matching result is the database image with the highest score. Finally, the matching score of the latent pair is evaluated by windowed uniqueness thresholding [37] to remove low-confidence matching results. The sequence matching parameters follow those optimized parameters in [34]. The parameters of cone region boundaries $v_{min}$ and $v_{max}$ are set as 0.4 and 2.5 respectively. The length of the matching sequence $n_q$ is set as 10, and the window width of uniqueness thresholding is also set as 10.

## IV. EXPERIMENTS

In this section, we present the dataset configurations, the training procedures of the unified network and the experimental results of assistive navigation.

### A. TRAINING OF THE UNIFIED NETWORK

The training of the unified network was carried on NVIDIA GeForce RTX 2080 Ti.

#### 1) SCENE RECOGNITION

The scene recognition branch is composed of the backbone part (BaseNet and Branch-1), pooling and fully connected layers. All of the layers in the branch are trained for scene recognition. The input size of images is set to 224 × 224. Different data augmentation approaches are applied to the training images, including random cropping and random flip. Cross entropy loss serves as the loss function during training.

##### *a: DATASET*

In the first training phase, the branch of scene recognition is trained on Places-365 dataset, which is a widely-used dataset for scene classification and scene features extraction. There are 1.8 million training images from 365 scene categories. Moreover, there are also 50 images per category in the validation set. In Figure 5, some instances of dataset images reveal that the dataset features a broad scale of real-world scene categories.

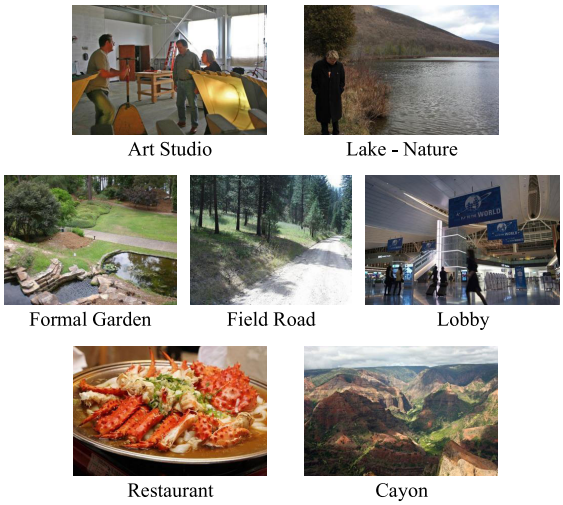


**FIGURE 5.** Several image instances with scene label in Places-365.

##### *b: PARAMETERS*

During training, SGD (stochastic gradient descending) is leveraged as the optimizer of the scene recognition branch. The learning rate is set as 0.01, the momentum is set as 0.9, and the weight decay is set as $10^{-4}$. The learning rate decays to 0.1 times of original value for every 30 epochs. The training of the scene recognition branch starts from the pre-trained results trained by the other task, so we use a larger learning rate to reach the optimal region of parameter space faster.

##### *c: METRICS*

The top-1 precision and top-5 precision, denoted as P@1 and P@5 respectively, are used as the performance criterion of scene recognition. If one of the top-K predictions hits the ground truth, the prediction is defined as a true positive. Then, top-K precision is defined as the number ratio of true positives to all of the testing samples.

##### *d: RESULTS*

In Table 4, we list the scene recognition performance of different convolutional networks on the validation set of Places-365. The computational complexity criterion (FLOPs) is given when the input images are with the size of 224 × 224 × 3. The trained models of other networks (i.e., VGGNet, GoogLeNet and ResNets) are provided by Zhou *et al.* [4]. It is clear that the classification performance of the two efficient models surpasses that of Resnet-18. Although the efficient networks achieve lower precision compared with other baselines, the computational efficiency of the two efficient networks is much more superior, considering they feature much fewer FLOPs. As for the two efficient models, ShuffleNet V2 is superior to MobileNet V2 in performance slightly.

**TABLE 4.** The performance of the scene recognition branch.

| Network | Complexity (FLOPs) | P@1 | P@5 |
|---|---|---|---|
| MobileNet V2 | 319M | 49.10% | 80.05% |
| ShuffleNet V2 | 150M | 52.35% | 82.92% |
| VGG-16 [4] | 15.5G | 55.24% | 84.91% |
| GoogLeNet [4] | 1.51G | 53.63% | 83.88% |
| ResNet-18 [4] | 1.82G | 26.64% | 54.96% |
| ResNet-50 [4] | 4.12G | 54.77% | 84.93% |

#### 2) SCENE DESCRIPTION

As shown in Figure 1, the scene description branch is composed of the backbone part (BaseNet and Branch-2) and the NetVLAD module. In the proposed unified network, BaseNet is the common part that is shared with the scene recognition branch. The initial weights of the scene description branch are set as the trained parameters of the scene recognition branch. Different from the case of scene recognition training, parameters in BaseNet are fixed, and only the layers in Branch-2 and NetVLAD are trained for scene description. Naturally, the number of trainable layers is adjustable, which means that the separation point of BaseNet and Branch-1/Branch-2 can be at different positions in the backbone

network. The weakly-supervised triplet ranking loss [5] is used to train the scene description branch.

#### a: DATASET

The branch of scene description is subsequently trained on Pittsburgh dataset [5]. As shown in Table 5, the dataset is composed of training subset, validation subset and testing subset. The discriminate ability of NetVLAD descriptors is obtained from the training phase, when the images with diverse appearances (e.g. different illumination, viewpoint and dynamic objects) but captured at the same place are leveraged as training data.

**TABLE 5. The number of images of the Pittsburgh dataset.**

| Dataset | Query Set | Database Set |
|---|---|---|
| Training Set | 7,416 | 10,000 |
| Validation Set | 7,608 | 10,000 |
| Testing Set | 6,818 | 10,000 |

#### b: PARAMETERS

Same with the other branch, SGD is also leveraged as the optimizer of this branch. The learning rate is set as $10^{-4}$, the momentum is set as 0.9, and the weight decay is set as $10^{-3}$. The learning rate decays to 0.5 times of original value for every 5 epochs. The scene description branch is fine-tuned based on the trained parameters derived from scene recognition, so we use a smaller learning rate.

#### c: METRICS

Faiss [38] is used to retrieve the nearest database descriptors for each query descriptor. If one of the retrieved top-K results falls within the range (25m) of the query image, it is defined as a correct retrieval. The ratio of correct retrievals to all the retrievals is denoted as precision.

#### d: RESULTS

In Table 6, the retrieval precisions at top-1, top-5 and top-10 of NetVLAD based on different backbones with different Branch-2 volume are presented. In the table, the Branch-2 volume is actually the number of blocks in Branch-2 for MobileNet V2 and ShuffleNet V2. That is to say, the trained $n$ layers are the last $n$ blocks of the backbone shown in Figure 2 and Table 2. For the case of Wide ResNet-18, the volume denotes the number of block groups that are defined in [39]. For VGG-based NetVLAD, the Branch-2 volume denotes the number of convolutional (pooling) layers, and is set as 5, which is the optimal training configuration presented in [5].

Obviously, the high-complexity networks (Wide ResNet-18 and VGG-16) yield higher precision. Fortunately, the network based on MobileNet V2 achieves the precision that is close to those high-complexity networks and is acceptable. It is worthwhile to note that ShuffleNet V2 is inferior to MobileNet V2 on the task of scene description although the last feature maps of the ShuffleNet V2 backbone have deeper channels (464 dimensions) than that of the MobileNet V2 backbone (320 dimensions). Considering that ShuffleNet V2 performs better than MobileNet V2 on scene recognition, it reveals that the network performing superior on scene recognition does not necessarily feature the superior performance on scene description. In view of the balance between precision and efficiency, we choose MobileNet V2 with 11 trained layers as the optimal configuration and use it in the following experiments.

Furthermore, we conducted an ablation study on the performance boost of scene description by training from scene recognition models. In this work, the joint training of scene recognition and scene description is actually the common part of two network branches, i.e. BaseNet. Therefore, the separate training means that the BaseNet part of NetVLAD network has nothing to do with the scene recognition training results. Therefore, the branch of scene description is trained based on the classification results on ImageNet dataset [40], and the training parameters are the same with those mentioned above. The image retrieval performance is also presented in Table 6, from which we can see that the NetVLAD descriptor based on ImageNet is obviously inferior to that trained based on the scene recognition branch. Thereby, training firstly on scene recognition does improve the performance of scene description.

The final scene descriptor features the dimensions of $K \times D$, and the dimension for MobileNet V2-based NetVLAD is $20,480$, which is large for image retrieval. The extracted descriptor is compressed to a lower scale for computational efficiency. Therefore, PCA (principal component analysis) with whitening [41] and L2-normalization are used to compress the dimensions of global descriptor and to speed up image retrieval.

The description performance of different reduced dimensions are tested on the Pittsburgh dataset, and the results are presented in Table 7. For each kind of compressed descriptor, the retrieval time per query on Pittsburgh-Test set is also presented. According to the results, the NetVLAD descriptors should be compressed to $2,560$ dimensions, where the top-1 precision drops only by about 3% (we think it is acceptable) compared with the original descriptors and the descriptor matching time is dramatically reduced. If we continue to compress dimensions, the matching time is no longer reduced significantly, and the recall is reduced largely. Therefore, we find a balanced trade-off between the descriptor matching time and localization precision at $2,560$ dimensions.

### B. ASSISTIVE DEVICES AND REAL-WORLD DATASETS

We have developed a wearable assistive device Intoer [42] for navigational assistance. The device has been used in different assistive applications, including obstacle avoidance [12], [43], traffic intersection assistance [13]–[15], and visual localization [33], [34], [44], etc. As shown in Figure 6, Intoer is composed of the multi-modal camera RealSense [45], a customized portable processor with GNSS module, and a pair of bone-conduction earphones [46].

**TABLE 6.** The performance of the scene description branch.

| Network | Branch-2 Volume | Pitts-Val | | | Pitts-Test | | |
|---|---|---|---|---|---|---|---|
| | | P@1 | P@5 | P@10 | P@1 | P@5 | P@10 |
| MobileNet V2 | 5 | 74.86% | 90.89% | 94.57% | 74.22% | 88.22% | 91.97% |
| | 7 | 76.14% | 91.73% | 95.04% | 75.19% | 88.56% | 91.90% |
| | 11 | 78.84% | 92.84% | 95.69% | 76.51% | 89.57% | 92.93% |
| | 14 | 80.86% | 93.55% | 95.99% | 78.23% | 90.20% | 93.34% |
| MobileNet V2* | 7 | 71.44% | 87.79% | 92.11% | 69.20% | 85.34% | 89.88% |
| | 11 | 76.29% | 90.34% | 93.60% | 73.15% | 87.22% | 91.53% |
| | 14 | 73.50% | 88.01% | 91.93% | 71.90% | 87.57% | 91.67% |
| ShuffleNet V2 | 5 | 60.08% | 80.13% | 85.65% | 57.35% | 77.11% | 84.51% |
| | 7 | 63.51% | 82.02% | 87.00% | 61.66% | 79.75% | 88.58% |
| | 11 | 53.08% | 73.80% | 80.02% | 55.91% | 76.82% | 83.86% |
| Wide ResNet-18 | 1 | 76.27% | 89.02% | 92.69% | 71.76% | 86.55% | 90.54% |
| | 2 | 83.71% | 94.26% | 96.46% | 81.68% | 90.36% | 93.32% |
| VGG-16 | 5 | 85.27% | 94.68% | 97.00% | 81.90% | 91.23% | 95.75% |

* denotes that the training of the scene description branch is based on the pre-trained model of ImageNet dataset.

**TABLE 7.** The dimension reduction performance of NetVLAD descriptor.

| Dimensions | Time (ms) | Pittsburgh-Validation | | | Pittsburgh-Test | | |
|---|---|---|---|---|---|---|---|
| | | P@1 | P@5 | P@10 | P@1 | P@5 | P@10 |
| 20,480 | 106.33 | 78.84% | 92.84% | 95.69% | 76.51% | 89.57% | 92.93% |
| 5,120 | 1.14 | 77.20% | 92.07% | 95.10% | 74.57% | 88.37% | 92.50% |
| 2,560 | 0.62 | 75.76% | 91.13% | 94.62% | 73.42% | 88.00% | 92.28% |
| 1,280 | 0.31 | 71.12% | 88.54% | 92.67% | 69.84% | 86.25% | 91.12% |
| 640 | 0.16 | 58.18% | 80.40% | 87.70% | 59.86% | 80.85% | 87.41% |
| 320 | 0.08 | 30.49% | 57.32% | 68.95% | 32.61% | 59.51% | 71.60% |

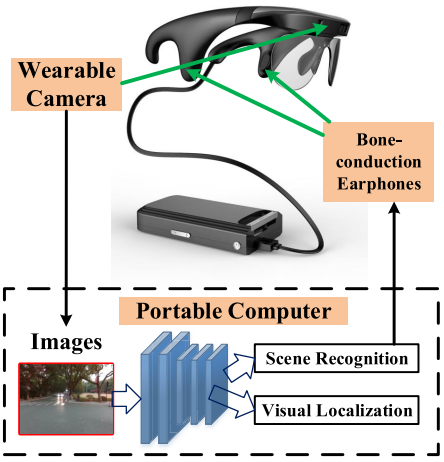


**FIGURE 6.** The assistive device Intoer is used to achieve the proposed assistive navigation pipeline.

**TABLE 8.** The detailed information of the west lake dataset.

| Route | # Database | # Query |
|---|---|---|
| Red | 44 | 44 |
| Yellow | 58 | 50 |
| Blue | 54 | 85 |

In this paper, we use Intoer to capture image datasets for visual localization and scene recognition. The West Lake dataset was collected in the scenic area of the West Lake, and was previously leveraged to validate Visual Localizer [44] qualitatively. As shown in Table 8, this dataset includes three routes, on which the camera records images with an interval of three seconds. In view of the walking speed, the interval distance of consecutive images is estimated as 3-4 meters. We labeled each query image with several scene classes and no more than one corresponding database image, so as to evaluate the proposed pipeline quantitatively. Apart from that, we also labeled each database image with several scene classes. In view of the semantic ambiguity of the scene recognition, the ground truths of each image are multiple scene categories.

The route schematic of the West Lake dataset is shown in Figure 7. It is worth noting that the query images and the database images were captured in a winter afternoon and a summer morning respectively, which results in the illumination and vegetation variations. Moreover, the dataset was collected in the open area (the West Lake scenic area) where pedestrians occurred in images frequently and object layouts also vary between the query and the database. Thereby, the dynamic objects compose one portion of real-world scenarios and interfere with the stability of visual localization. As images are captured by the wearable camera, the viewpoint of the query and that of the database is different and some images suffer from motion blur. In summary, there are significant appearance variations between the queries and the databases in the real-world dataset, which is important to validate the robustness of visual localization.

### C. EXPERIMENTAL RESULTS AND VISUALIZATION

In this section, the experimental results of the proposed pipeline on the real-world dataset are presented and analyzed. Firstly, the performance test of scene recognition is carried out on the West Lake dataset. Then, we evaluate the utility


**FIGURE 7.** The route schematic and some instances of the West Lake dataset.

**TABLE 9.** The scene recognition performance of the west lake dataset.

| Route | P@1 | P@3 | P@5 |
|---|---|---|---|
| West Lake - Red | 70.45% | 86.36% | 88.64% |
| West Lake - Yellow | 75.86% | 82.76% | 89.66% |
| West Lake - Blue | 64.81% | 77.78% | 81.48% |

of visual localization on the West Lake dataset and Gardens Point Walking dataset [47].

As mentioned before, MobileNet V2 is chosen as the backbone of the proposed network. For both branches of the unified network, the layer volume of Branch-1 / Branch-2 is 11, and the NetVLAD descriptors are compressed to 2,560 dimensions. All of the input images are resized to 224 (smaller edge) with the respect ratio invariant, and then cropped to the size of $224 \times 224$.

### 1) SCENE RECOGNITION

In view of the scenes overlap between database images and query images, we performed the scene recognition experiment on the database portion of the West Lake dataset. As mentioned before, top-K precision is used as the accuracy indicator of scene recognition. The scene recognition precision on the West Lake dataset is presented as Table 9.

As we can see, the scene recognition of the unified network achieves satisfactory performance in the real-world assistive scenarios, though the network is only trained on the public Places dataset. On the challenging dataset, the top-5 precision of scene recognition is more than 80%. Some top-1 predictions of scene recognition are presented in Figure 8. Fortunately, it is concluded from the results that some negative predictions (i.e. Yellow-40 and Yellow-50 in Figure 8) are not irrelevant to the actual scenes in images.

### 2) VISUAL LOCALIZATION

To quantitatively evaluate the performance of visual localization, we use a more rigorous criterion than top-K precision used in the previous scene description validation. Considering the query and database images are sequential in the dataset, the localization result of a query image is represented as the sequence index of the best-matching database image. If the index difference between the place recognition result and the ground truth (denoted as *Error*) is less than or equal to the tolerance (denoted as *tol*), the result is correct. Based on that, the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) are defined in Table 10. Precision is defined as the proportion of true positives out of all predicted positives, and recall is defined as the proportion of true positives to all of the ground-truth positives. Based on precision and recall, $F_1$ score is defined as

$$F_1 = 2 \times \frac{Recall \times Precision}{Recall + Precision}. \tag{6}$$

Using different windowed uniqueness thresholds, a series of precision and recall is obtained and plotted in Figure 9. Generally speaking, the utility of visual localization achieves superior performance under the circumstances of assistive navigation. The yellow route of the dataset is relatively easy, because the places in the yellow route have more diverse appearances than the other routes. Therefore, the performance of the yellow route is perfect, and is the best among the three routes. There are plenty of aliasing places, which resemble each other in terms of visual appearance but differ in terms of spatial position existing in the red and blue route. Moreover, the two routes contain the new queries that do not appear in database. On those hard scenarios, the performance of the proposed pipeline is also acceptable.

In order to inspect visual localization performance deeply, we visualize those image matching results by assigning the window uniqueness threshold to be 1.2, at which the precision is relatively high meanwhile the recall does not drop heavily. For all of the three routes, the visual localization precision is over 60% under a recall of 60%. In this condition, the visualization results of visual localization are shown in Figure 10,


**FIGURE 8.** The top-1 predictions of the West Lake dataset. Green denotes correct result, and red denotes incorrect result.

**TABLE 10.** The definitions of prediction attribute for visual localization.

| | $Error < tol$ | Ground Truth | |
|---|---|---|---|
| | | Positive | Negative |
| Predictions | Yes | TP | TN |
| | No (Positive) | FP | FP |
| | No (Negative) | FN | - |

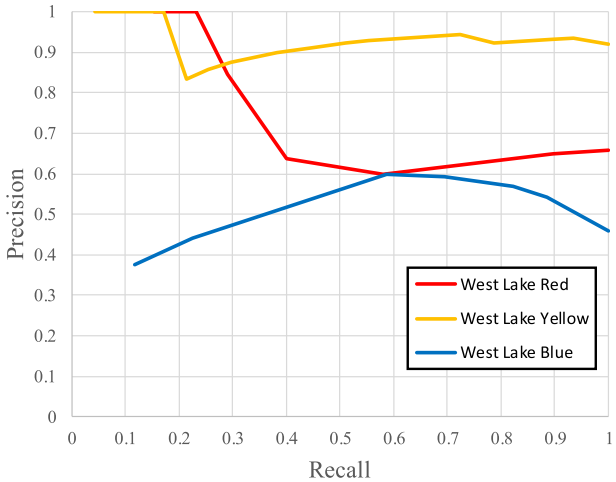


**FIGURE 9.** The precision-recall curve of visual localization on the West Lake dataset.

where the images are sampled uniformly from the query sequence. As we mention ahead, the localization results of yellow routes match well also in terms of visual intuition. As for the other hard routes, some localization failures are caused by false negative predictions, such as query 10 and query 20 in the red route as well as query 30 and query 50 in the blue route. The false negatives result in less hazardous than false positives in localization applications.

In order to illustrate the generalization of the proposed pipeline, we carried out the experiment on the Gardens Point Walking dataset, which was collected in a campus using a portable camera. The scenario of the public dataset is close to the assistive navigation, though it is less challenging compared with our West Lake dataset. We choose ''day-left'' and ''day-right''sequence as the query and the database respectively. In Table 11, we present the visual localization performance of three approaches: the proposed pipeline, the approach using an off-the-shelf AlexNet layer as the image descriptor [47], and multi-level algorithm [48], which uses global descriptors (GoogLeNet [18]) and landmark-based local descriptors (GeoDesc [49]) to achieve coarse-to-fine visual localization. For the multi-level algorithm, we choose the configuration with the best localization performance, i.e. using Fundamental Matrix to verify geometric transformation. Similarly, we choose the optimal configuration (using FC6 as the descriptor) of Sünderhauf's method [47] as the comparison baseline of our method.

Despite the fact that we use a much simpler backbone network and do not use local feature-based geometric verification at all, the proposed pipeline achieves superior performance when $tol = 5$. The index difference tolerances of 3 and 5 correspond to around 10 meters and 18 meters in real-world space respectively. It is worthwhile to note that Sünderhauf's method achieves the optimal performance when using Conv3 layer on any other datasets except Gardens Point Walking dataset [47]. Under that configuration,


**FIGURE 10.** The visualization of some results in visual localization on (a) the red route, (b) the yellow route, and (c) the blue route of the West Lake dataset. The correct matching results are denoted with green, and the incorrect ones are denoted as red.

**TABLE 11.** The visual localization performance of the gardens point dataset.

| Approaches | $tol = 3$ | | | $tol = 5$ | | |
|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ |
| Sünderhauf's [47] | - | - | - | - | - | **0.96** |
| Multi-level [48] | 81.5% | 100% | **0.90** | 89.5% | 100% | 0.94 |
| Ours | 77.5% | 99.0% | 0.87 | 93.5% | 99.0% | **0.96** |

Sünderhauf's method yields a $F_1$ score of 0.89, which illustrates that the proposed pipeline outperforms Sünderhauf's method in terms of performance stability against algorithm configurations.

Finally, we discuss computational efficiency performance of the proposed pipeline for assistive navigation. As presented in [36], the inference speed of the backbone network (MobileNet V2) used in this paper achieves 8.9 frames/second on ARM processors. Meanwhile, we have previously presented in [34] that the sequence matching achieves a speed of around 40 frames/second on a portable x86 CPU. Considering that the image sampling interval of the dataset is 3 seconds, the proposed pipeline is able to run in real time.

## V. CONCLUSION

Aiming to resolve the two key issues of assistive navigation "Where am I?" and "What are the surroundings?", we propose a new unified network combining scene description and scene recognition in this paper. The unified network is based on the compact convolutional networks MobileNet V2. The proposed pipeline leverages image retrieval and sequence matching to yield the final results of visual localization. During the training phase, the network configurations are tuned on the public datasets to search for the optimal network structure.

In view of the challenging scenario of assistive navigation, the real-world datasets are proposed to evaluate the proposed pipeline comprehensively. The proposed assistive navigation pipeline is proved to achieve satisfactory performance on the real-world datasets. The top-5 precision of scene recognition is more than 80%, and the visual localization precision is over 60% under a recall of 60%.

In the future, we plan to extend this work to 6-DoF localization based on the unified network, where both the global descriptors and the local descriptors are implemented in a single deep network.

## REFERENCES

[1] W. Elmannai and K. Elleithy, "Sensor-based assistive devices for visually-impaired people: Current status, challenges, and future directions," *Sensors*, vol. 17, no. 3, p. 565, Mar. 2017.
[2] R. Tapu, B. Mocanu, and T. Zaharia, "Wearable assistive devices for visually impaired: A state of the art survey," *Pattern Recognit. Lett.*, early access, Oct. 29, 2018, doi: 10.1016/j.patrec.2018.10.031.
[3] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.
[4] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[5] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, Jun. 2018.

[6] J. Q. Peng, Y. H. Liu, C. Y. Lyu, Y. H. Li, W. G. Zhou, and K. Fan, "FPGA-based parallel hardware architecture for SIFT algorithm," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot. (RCAR)*, Jun. 2016, pp. 277–282.

[7] J. Peng, W. Xu, and H. Yuan, "An efficient pose measurement method of a space non-cooperative target based on stereo vision," *IEEE Access*, vol. 5, pp. 22344–22362, 2017.

[8] L. Grewe, A. Kashyap, K. Chandran, A. Shahshahani, and J. Shahshahani, "iSight: Computer vision based system to assist low vision," *Proc. SPIE*, vol. 10646, pp. 242–249, Apr. 2018.

[9] A. Perez-Yus, D. Gutierrez-Gomez, G. Lopez-Nicolas, and J. J. Guerrero, "Stairs detection with odometry-aided traversal from a wearable RGB-D camera," *Comput. Vis. Image Understand.*, vol. 154, pp. 192–205, Jan. 2017.

[10] S. Lin, K. Wang, K. Yang, and R. Cheng, "KrNet: A kinetic real-time convolutional neural network for navigational assistance," in *Proc. Int. Conf. Comput. Helping People With Special Needs (ICCHP)*, 2018, pp. 55–62.

[11] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "A compressive sensing approach to describe indoor scenes for blind people," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 7, pp. 1246–1257, Jul. 2015.

[12] K. Yang, K. Wang, L. Bergasa, E. Romera, W. Hu, D. Sun, J. Sun, R. Cheng, T. Chen, and E. López, "Unifying terrain awareness for the visually impaired through real-time semantic segmentation," *Sensors*, vol. 18, no. 5, p. 1506, May 2018.

[13] R. Cheng, K. Wang, and S. Lin, "Intersection navigation for people with visual impairment," in *Proc. Int. Conf. Comput. Helping People With Special Needs (ICCHP)*, 2018, pp. 78–85.

[14] R. Cheng, K. Wang, K. Yang, N. Long, and W. Hu, "Crosswalk navigation for people with visual impairments on a wearable device," *J. Electron. Imag.*, vol. 26, no. 5, pp. 1–14, Oct. 2017.

[15] R. Cheng, K. Wang, K. Yang, N. Long, J. Bai, and D. Liu, "Real-time pedestrian crossing lights detection algorithm for the visually impaired," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 20651–20671, Aug. 2018.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[20] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: https://arxiv.org/abs/1602.07360

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[22] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[24] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Robot. Auton. Syst.*, vol. 69, pp. 15–27, Jul. 2015.

[25] E. Pepperell, P. Corke, and M. Milford, "Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints," *Int. J. Robot. Res.*, vol. 35, no. 9, pp. 1057–1179, Aug. 2016.

[26] M. Abdollahyan, S. Cascianelli, E. Bellocchio, G. Costante, T. A. Ciarfuglia, F. Bianconi, F. Smeraldi, and M. L. Fravolini, "Visual localization in the presence of appearance changes using the partial order kernel," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 697–701.

[27] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," in *Proc. Vis. Place Recognit. Changing Environ. Workshop, IEEE Int. Conf. Robot. Autom. (ICRA)*, Hong Kong, vol. 2, 2014, p. 3.

[28] T. A. Ciarfuglia, G. Costante, P. Valigi, and E. Ricci, "A discriminative approach for appearance based loop closing," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 3837–3843.

[29] E. S. Stumm, C. Mei, and S. Lacroix, "Building location models for visual place recognition," *Int. J. Robot. Res.*, vol. 35, no. 4, pp. 334–356, Apr. 2016.

[30] S. Cascianelli, G. Costante, E. Bellocchio, P. Valigi, M. L. Fravolini, and T. A. Ciarfuglia, "Robust visual semi-semantic loop closure detection by a covisibility graph and CNN features," *Robot. Auton. Syst.*, vol. 92, pp. 53–65, Jun. 2017.

[31] J.-E. Kim, M. Bessho, S. Kobayashi, N. Koshizuka, and K. Sakamura, "Navigating visually impaired travelers in a large train station using smartphone and Bluetooth low energy," in *Proc. 31st Annu. ACM Symp. Appl. Comput. (SAC)*, 2016, pp. 604–611.

[32] C. Gleason, D. Ahmetovic, S. Savage, C. Toxtli, C. Posthuma, C. Asakawa, K. M. Kitani, and J. P. Bigham, "Crowdsourcing the installation and maintenance of indoor localization infrastructure to support blind navigation," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 9:1–9:25, Mar. 2018.

[33] R. Cheng, K. Wang, L. Lin, and K. Yang, "Visual localization of key positions for visually impaired people," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2893–2898.

[34] R. Cheng, K. Wang, J. Bai, and Z. Xu, "OpenMPR: Recognize places using multimodal data for people with visual impairments," *Meas. Sci. Technol.*, vol. 30, no. 12, Dec. 2019, Art. no. 124004.

[35] W. Hu, K. Wang, H. Chen, R. Cheng, and K. Yang, "An indoor positioning framework based on panoramic visual odometry for visually impaired people," *Meas. Sci. Technol.*, vol. 31, no. 1, Jan. 2020, Art. no. 014006.

[36] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Computer Vision*. Cham, Switzerland: Springer, 2018, pp. 122–138.

[37] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2012, pp. 1643–1649.

[38] J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, early access, Jun. 7, 2019, doi: 10.1109/TBDATA.2019.2921572.

[39] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, R. C. Wilson, E. R. Hancock, and W. A. P. Smith, Eds. London, U.K.: BMVA Press, Sep. 2016, pp. 87.1–87.12, doi: 10.5244/C.30.87.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[41] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening," in *Computer Vision*. Berlin, Germany: Springer, 2012, pp. 774–787.

[42] KrVision. (2017). *Intoer: Auxiliary Glasses for People With Visual Impairments*. (in Chinese). [Online]. Available: http://www.krvision.cn/cpjs/

[43] K. Yang, K. Wang, R. Cheng, W. Hu, X. Huang, and J. Bai, "Detecting traversable area and water hazards for the visually impaired with a pRGB-D sensor," *Sensors*, vol. 17, no. 8, p. 1890, 2017.

[44] S. Lin, R. Cheng, K. Wang, and K. Yang, "Visual localizer: Outdoor localization based on ConvNet descriptor and global optimization for visually impaired pedestrians," *Sensors*, vol. 18, no. 8, p. 2476, 2018.

[45] Intel. (2017). *Realsense Camera ZR300*. [Online]. Available: https://software.intel.com/en-us/realsense/zr300

[46] AfterShokz. (2018). *Bone-Conduction Earphone*. [Online]. Available: https://www.aftershokz.com/

[47] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 4297–4304.

[48] Y. Fang, K. Wang, R. Cheng, K. Yang, and J. Bai, "Visual place recognition based on multilevel descriptors for the visually impaired people," *Proc. SPIE*, vol. 11158, Oct. 2019, Art. no. 1115808.

[49] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan, "GeoDesc: Learning local descriptors by integrating geometry constraints," in *Computer Vision*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 170–185.

**RUIQI CHENG** was born in Weihai, Shandong, China, in 1992. He received the bachelor's degree in information engineering from Zhejiang University, China, where he is currently pursuing the Ph.D. degree with the College of Optical Science and Engineering.

He is the author of more than 20 articles. He holds more than ten patents. His research interests include computer vision and image sensing for assistive navigation and autonomous vehicles.

**KAIWEI WANG** received the B.S. and Ph.D. degrees from Tsinghua University, China, in 2001 and 2005, respectively.

From 2005 to 2009, he was with the Center of Precision Technologies, University of Huddersfield, U.K. Since 2009, he has been with Zhejiang University, China, where he is currently the Deputy Director of the National Engineering Research Center of Optical Instrumentation and a Professor with the College of Optical Science and Engineering. He has published more than 150 research publications in international journals and professional conferences. His main research interests include optical sensing and measurement and assistive technology for visually impaired people.

**JIAN BAI** was born in 1967. He received the Ph.D. degree from Zhejiang University, China. He is currently the Director of the Institute of Optical Engineering and a Professor with the College of Optical Science and Engineering, Zhejiang University.

He is the author of more than 100 articles. He holds over 20 Chinese National patents. His research interests include optical design and optical testing, especially refractive/diffractive optical imaging and panoramic annular imaging

**ZHIJIE XU** (Member, IEEE) received the B.Eng. degree in communication engineering from the Xi'an University of Science and Technology, in 1991, and the Ph.D. degree from the University of Derby, in 2000.

He holds a tenured academic postion at the University of Huddersfield. He has published over 100 peer-reviewed journal articles and conference papers and editing five books in the relevant fields. His main research interests include visual computing, vision systems, data science, and machine learning.

Dr. Xu is a member of IET, BCS, BMVA, and a fellow of HEA. He has been serving as an editor, a reviewer, and the chair for many prestigious academic journals and conferences.