Trans4Trans: Efficient Transformer for Transparent Object and Semantic Scene Segmentation in Real-World Navigation Assistance

Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen, *Member, IEEE*

***Abstract*— Transparent objects, such as glass walls and doors, constitute architectural obstacles hindering the mobility of people with low vision or blindness. For instance, the open space behind glass doors is inaccessible, unless it is correctly perceived and interacted with. However, traditional assistive technologies rarely cover the segmentation of these safety-critical transparent objects. In this paper, we build a wearable system with a novel dual-head Transformer for Transparency *(Trans4Trans)* perception model, which can segment general- and transparent objects. The two dense segmentation results are further combined with depth information in the system to help users navigate safely and assist them to negotiate transparent obstacles. We propose a lightweight Transformer Parsing Module *(TPM)* to perform multi-scale feature interpretation in the transformer-based decoder. Benefiting from TPM, the double decoders can perform joint learning from corresponding datasets to pursue robustness, meanwhile maintain efficiency on a portable GPU, with negligible calculation increase. The entire Trans4Trans model is constructed in a symmetrical encoder-decoder architecture, which outperforms state-of-the-art methods on the test sets of Stanford2D3D and Trans10K-v2 datasets, obtaining mIoU of 45.13% and 75.14%, respectively. Through a user study and various pre-tests conducted in indoor and outdoor scenes, the usability and reliability of our assistive system have been extensively verified. Meanwhile, the Tran4Trans model has outstanding performances on driving scene datasets. On Cityscapes, ACDC, and DADA-seg datasets corresponding to common environments, adverse weather, and traffic accident scenarios, mIoU scores of 81.5%, 76.3%, and 39.2% are obtained, demonstrating its high efficiency and robustness for real-world transportation applications.**

***Index Terms*— Computer vision for the visually impaired, wearable assistive system, transparent object segmentation, semantic segmentation, scene understanding.**

Manuscript received 20 August 2021; revised 31 December 2021 and 1 March 2022; accepted 14 March 2022. Date of publication 28 March 2022; date of current version 11 October 2022. This work was supported in part by the Federal Ministry of Labor and Social Affairs (BMAS) through the Accessible Maps Project under Grant 01KM151112, in part by the University of Excellence through the "KIT Future Fields" Project, and in part by Hangzhou SurImage Company Ltd. The Associate Editor for this article was Z. Duric. *(Corresponding author: Kailun Yang.)*

Jiaming Zhang, Kailun Yang, and Kunyu Peng are with the Computer Vision for Human–Computer Interaction Laboratory, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany (e-mail: kailun.yang@kit.edu).

Angela Constantinescu and Karin Müller are with the Center for Digital Accessibility and Assistive Technology, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany.

Rainer Stiefelhagen is with the Computer Vision for Human–Computer Interaction Laboratory and the Center for Digital Accessibility and Assistive Technology, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany.

Code will be available at: github.com/jamycheung/Trans4Trans.

Digital Object Identifier 10.1109/TITS.2022.3161141

I. Introduction

ASSISTED navigation of pedestrians and automated driving of intelligent vehicles are inextricably intertwined in the Intelligent Transportation Systems (ITS) field [1]–[4], both with the aim to improve traffic flow towards the utopia of all road participants. In addition to vehicles from the driving perspective, humans and their mobilities from the walking perspective are involved. However, people with disabilities may have difficulties in using transportation infrastructures, and the bottleneck of inclusiveness should be broken in ITS. To this end, it is necessary to expand the coverage of assistance systems from drivers to pedestrians, especially those with visual impairments, who are one of the most vulnerable road users [5].

To assist the navigation of visually impaired people, it is essential to attain efficient and robust *walking* scene understanding, which shares similar challenges with the ITS research line on *driving* surrounding segmentation [6], [7]. In real-world scenes, modern fully glazed facades and transparent objects are very common, but they are rarely addressed in existing semantic perception systems either for automated transportation or assisted navigation. Knowledge of glass architecture [8] and glass doors [9], [10] are particular important for visually impaired people, because transparent objects often present architectural barriers which hinder the mobility of people with low vision or blindness. For example, if an inaccessible area blocked by a transparent door (Fig. 1(b)) is detected by an assistive system as accessible, it could lead to a wrong interaction and cause harm to the user.

However, understanding transparent objects is a puzzle for most vision-based navigation assistance systems [11], [12], as notoriously, satisfactory dense prediction of transparent objects is difficult to obtain. The appearance difference between glass doors and large glass windows is insignificant, thus it troubles people with residual sight [13]. Moreover, a system to identify landmarks such as doors is particularly appreciated by people with visual impairments, since finding a door or a building entrance is difficult due to the inaccuracy of GPS [13].

To tackle these challenges, we put forward a wearable system for walking scene perception, covering object- and walkable area segmentation, with the ultimate goal to enable

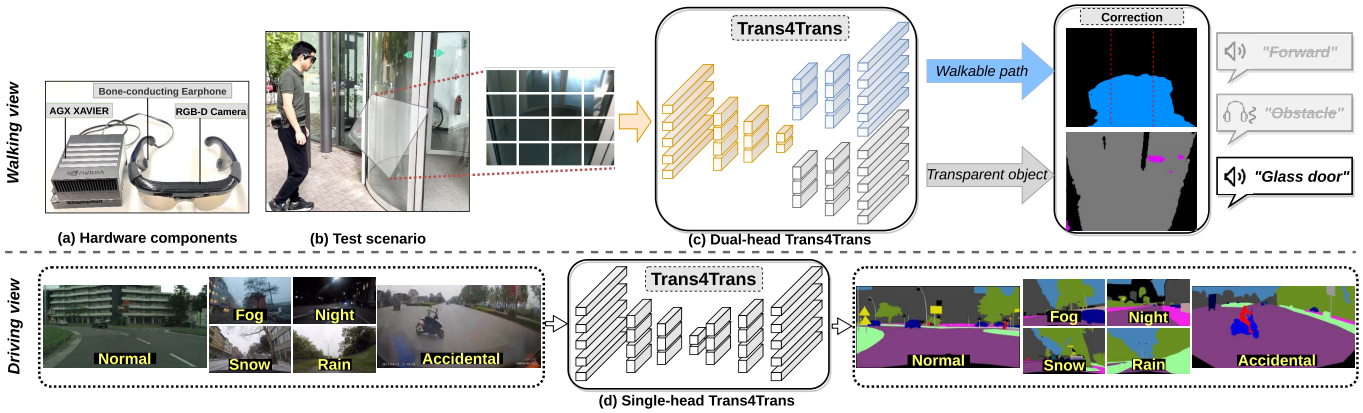


Fig. 1. (a) The system equipped with smart vision glasses and a portable processor is tested (b) in front of a glass door. The input image is segmented as ***walkable path*** and ***glass door*** by (c) our Trans4Trans model for navigation. The user interface has vibration and voice feedback. After training on normal and adverse data, (d) Trans4Trans reaches high robustness in various real-world driving scenarios.

visually impaired people navigate in real-world scenes safely and independently. We propose an efficient transformer-based semantic segmentation architecture dubbed *Trans4Trans*, precisely *Transformer for Transparency*, as shown in Fig. 1(c). Since transparent objects are often texture-less or share identical content as their surroundings, it is essential to associate long-range dependencies to robustly infer transparent regions. For this reason, Trans4Trans is established with both transformer-based encoder and decoder to fully exploit the long-range context modeling capacity of self-attention layers in transformers [14]. We design a *Transformer Parsing Module (TPM)* to fuse multi-scale feature maps generated from embeddings of dense partitions. The symmetric transformer-based decoder in a dual-head structure, thereby, can consistently parse features from the transformer-based encoder (see Fig. 1(c)). Along with semantically predicting general thing- and stuff categories like walkable areas, the dual-head design allows to segment transparent objects completely and accurately, which unifies various detection tasks and enhances the perception reliability relevant for safety-critical navigation assistance.

Trans4Trans has been integrated in our wearable assistive system (see Fig. 1(a)) comprising a pair of smart vision glasses and a mobile GPU processor. Based on the entire semantic information, the user interface is designed with a customized set of acoustic feedback, including the sonification of detected objects, the identified walkable directions, and warnings of close-range obstacles. The voice user interface yields intuitive navigation suggestions while requiring no prior knowledge.

Extensive experiments have been conducted on multiple datasets [15], [16] to verify the effectiveness of the assistive system. Particularly, the proposed model achieves state-of-the-art accuracy on the test sets of Stanford2D3D [15] and Trans10K-v2 [16]. Considering the synergy towards traffic safety and the shared challenges between walking- and driving scene understanding, Trans4Trans is further verified on driving benchmarks including Cityscapes [17], ACDC [18], and DADA-seg [19]. Trans4Trans surpasses HRNet [20] by $>20\%$ and $>12\%$ on ACDC and DADA-seg datasets with only its 20% GFLOPs, demonstrating its efficiency and robustness for various ITS application scenarios (see Fig. 1(d)).

Finally, a user study with visually impaired people and a series of field tests validate the usability and reliability of our portable system for navigational perception and assistance in the real world.

In summary, we deliver the following contributions:

- We build a wearable assistive system with a pair of smart vision glasses and a portable GPU processor for aiding people with visual impairments during navigation.
- We put forward an efficient segmentation architecture with transformer-based encoder and decoder, namely Transformer for Transparency (*Trans4Trans*). Its dual-head design can unify general- and transparent object segmentation. We propose a Transformer Parsing Module (TPM) to harvest multi-scale feature representations generated from embeddings of dense partitions.
- Trans4Trans has surpassed state-of-the-art Convolutional Neural Networks (CNNs) and transformer-based methods on Stanford2D3D and Trans10K-v2 datasets, while maintaining high efficiency on the assistive system.
- We design an assistive algorithm based on multiple segmentation results and the depth information. We create a user interface via a customized set of acoustic feedback and conduct a user study and various field tests, evidencing the usability and reliability of the assistive system.

This work is the extension of our conference paper [21]. The extended contents include:

- We advocate addressing driving scene segmentation from a novel perspective that jointly considers normal, adverse, and accidental scenarios.
- Transferring from the normal (Cityscapes) dataset to the adverse (ACDC) and accidental (DADA-seg) datasets, Trans4Trans brings $>20\%$ and $>12\%$ improvements with only 20% computational cost of the previous best HRNet. Transferring from the normal and adverse datasets, the performance on DADA-seg is increased to 39.2%.
- A comparison of different versions of transformer-based backbones is additionally conducted and the comprehensive analysis of the model efficiency is advanced.

- Through extra segmentation comparison between six different image scales and analysis of feature visualizations, key insights of transparent object segmentation are given.
- The user study is enriched with more details and expert-level comments and suggestions are summarized. Other enhanced parts are related work reviews, more detailed hardware design- and algorithm description, and additional failure analysis on various driving scenes.

II. Related Work

A. Semantic Segmentation for Visual Assistance

Whereas traditional assistance systems rely on multiple monocular detectors and depth sensors [3], [11], [22] for perception tasks, semantic segmentation can solve navigational detection problems in a unified way and has been employed in visual assistance. As pioneer approaches, semantic paintbrush [23] was proposed as an augmented reality system for people with low vision, while semantic labeling was applied to the problem of navigation using prosthetic vision [24].

Yang *et al.* [4] put forward to seize pixel-wise semantic segmentation to unify terrain detection tasks covering traversability perception, stairs navigation, and water-hazards negotiation. In [25] and [26], instance-specific segmentation models were directly used for content-aware sensing. Liu *et al.* [25] proposed HIDA to enable holistic indoor understanding for detection and avoidance of obstacles based on 3D point cloud semantic instance segmentation. Dense semantic segmentation has also been leveraged to address intersection perception like the detection of crosswalks, sidewalks, and blind roads [1], [4], [27]. Given a single RGB image, the segmentation model can be used to predict the geometry of both visible and occluded traversable surfaces [28], which can provide potential navigable routes. Moreover, we observe the trend of using semantic segmentation in various navigation assistance platforms [12], [29]. Yet, both of these systems cannot resolve the problems emerged with transparent objects.

B. Transparent Object Sensing

Classical visual assistance systems [22], [30] resorted to multi-sensor fusion to overcome the difficulties in handling transparent obstacles like glass objects, French windows, and French doors by using ultrasonic sensors together with RGB-D cameras. Multimodal and multispectral information were also frequently used. Okazawa *et al.* [31] performed simultaneous recognition of ordinary non-transparent objects and transparent objects by utilizing the difference in transmission characteristics under multispectral scenes. Moreover, polarization cues [32] and reflection priors [33] were often explored for transparency perception. For instance, Xiang *et al.* [34] built a polarization-driven semantic segmentation architecture by bridging RGB and polarization dimensions dynamically using efficient attention connections, which considers the optical features of polarimetric information for robust representation of diverse materials and lifts the performance of classes with polarization properties like *glass*.

Recently, large-scale transparent object segmentation datasets emerge [10], [16], [31], [34], [35]. Mei *et al.* [10] constructed the glass detection dataset in daily-life scenes. Xie *et al.* [16], [35] built the Trans10K dataset and validated that while pure RGB-based transparent object segmentation is rather a largely unsolved task, it is potential for real-world usages with the increased data amount. This spurs the community to go beyond traditional perception paradigms relying on sensor fusion and develop novel methods addressing transparent object segmentation. AdaptiveASPP [36] was designed as an enhanced version of ASPP [37] to appropriately harvest rich features at multi-stage levels in joint segmentation and boundary predictions. EBLNet [38] integrated a point-based graph convolution module to model global shape representations. Whereas these methods have reached high accuracy on glass-like object detection, we aim for a both efficient and robust semantic segmentation desirable for real-world navigation assistance. We establish a transformer-based system to assist transparency perception.

C. Attention- and Transformer-Based Semantic Segmentation

Since Fully Convolutional Networks (FCNs) [39] achieved semantic segmentation end-to-end by viewing it as a dense-pixel classification task, attention-based methods [40], [41] develop upon this paradigm. Inspired by Vision Transformer [42] that utilizes transformer layers to sequences of image patches for visual recognition, SETR [43] directly appends upsampling and segmentation heads atop ViT, encoding long-range context information from the very first layer. Along this line, various transformer architectures for dense image segmentation appear [44], [45]. PVT [46], [47] and SegFormer [45] propose pyramid structures of vision transformers for collecting hierarchical feature representations. ECANet [48] and CSWin transformer [49] advocate performing self-attention in horizontal or vertical stripes to achieve powerful modeling capacity while lowering computation overheads, while different patch sizes are extensively investigated in TimeSformer [50].

In this research, we propose an efficient *Trans4Trans* framework with focus set on assisting navigation of visually impaired people in the wild. Differing from existing works that either stack attention layers [40] and encoder-decoder transformers on CNN backbones [16], or employing CNN-based decoders on top of transformer encoders [43], [46], in *Trans4Trans* both encoder and decoder are based on transformer, together with a novel Transformer Parsing Module inserted in the dual-head decoder, which unifies transparent object and semantic scene segmentation.

III. System Architecture

A. Trans4Trans

Benefiting from the Multi-Head Self-Attention (MHSA) structure, ViT [42] introduced the modeling ability in acquiring long-range dependencies. SETR [43] extended transformer to segmentation and validated on high-resolution inputs. To reduce the costly computation of MHSA, Swin [44] and CSWin [49] limited the token region of the full attention by using shift operations or cross-shaped windows. However, the global context is slowly covered by stacking a large number of

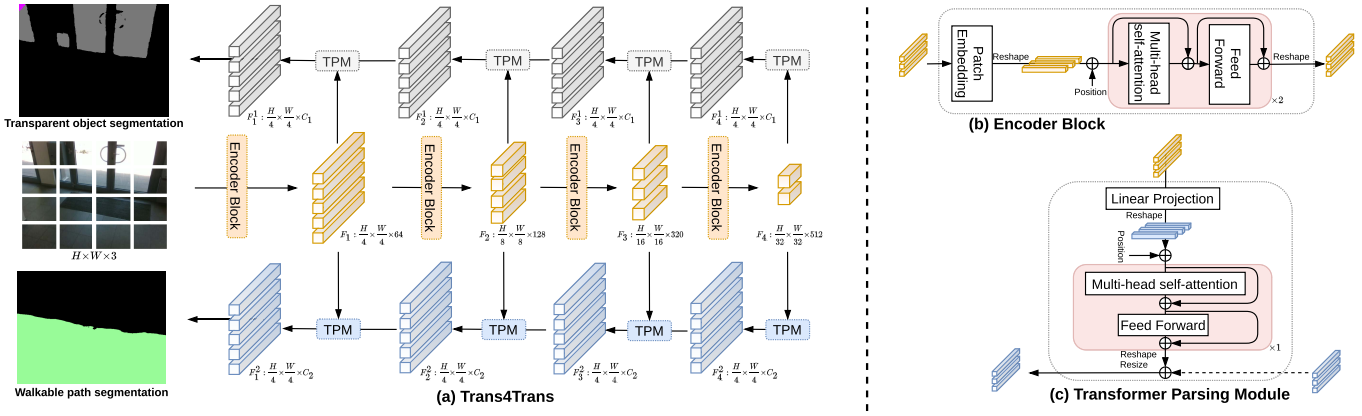


Fig. 2. The architecture of (a) Trans4Trans model consists of shared encoder and dual decoders, while (b) and (c) are the transformer-based encoder block and our proposed Transformer Parsing Module (TPM), respectively. The dashed arrow of TPM is absent in the last stage.

such regional self-attention blocks, and the receptive field is rather limited. In this work, we preserve the full-scale MHSA in the encoder for faster-enlarging receptive fields. To balance efficiency and accuracy in Trans4Trans, we consider: (1) to build upon a pyramidal transformer encoder with full-scale self-attention blocks; (2) to lighten the decoder to enable deploying a dual-head model on portable GPUs; and (3) to maintain a symmetrical encoder-decoder structure for obtaining pyramidal feature representations.

1) Pyramidal Feature Representations: The whole architecture of the Trans4Trans model is shown in Fig. 2(a). Similar to recently emerged models [45], [49], we split the image $H\times W\times 3$ with a 4×4 patch size, as a smaller patch size is more conducive for dense predictions [50]. A 7×7 convolution layer with a stride of 4 is utilized to perform overlapping patch embedding. Contrary to the single-scale feature of ViT, pyramidal features $\{F_1, F_2, F_3, F_4\}$ are obtained in our multi-scale encoder. Similar to the four-stage structure as ResNet [51], a 3×3 convolution layer with a stride of 2 is utilized within each stage of the encoder to progressively downsample the lower-level high-resolution features to the higher-level low-resolution features. In each stage, two full-scale MHSA blocks are stacked to realize faster-enlarging receptive fields. The encoder block is shown in Fig. 2(b), in which we use depth-wise convolutions in the Feed Forward module to reduce the computation. In specific, the pyramid features are extracted with $\{4, 8, 16, 32\}$ downsampling rates and $\{64, 128, 320, 512\}$ channels.

2) Transformer Parsing Module: Recent transformer methods [43], [44] opt to modify the encoder while neglecting the decoder design. Yet, a robust decoder is critical for deploying a transformer model in real applications, especially for an assistive system. The decoder is designed with four transformer-based stages corresponding to the multi-scale encoder, which is different to the single-scale decoder of ViT [42] and CNN-headed decoder of Trans2Seg [16]. To maintain the symmetrical pyramid features both in the encoder and the decoder, we propose the *Transformer Parsing Module (TPM)* to interpret the multi-scale features $\{F_1, F_2, F_3, F_4\}$, as illustrated in Fig. 2(a) and TPM is shown in detail in Fig. 2(c). A linear projection is used to translate the feature from the encoder into a preset embedding with a dimension C, which maintains the features channel-wise identical throughout the decoder. An MHSA module similar to the one in the encoder block is leveraged to obtain symmetrical features. An upsampling (resize) operation is used to preserve the same resolution $\frac{H}{4}\times\frac{W}{4}\times C$ of different features between adjacent stages. As a result in the hierarchical decoder, the lower-resolution features with long-range contextual information from higher stages are aggregated with the higher-resolution features with fine- and local information from shallow stages.

3) Multiple Decoders: Based on TPM, the decoder demands little computing resources and eases the integration with arbitrary encoder backbones, thus Trans4Trans is more flexible to be deployed on a portable system. To deal with the data-hunger problem of training transformers [42] and robustify segmentation in the wild, we explore a double-head design in transformer model to perform joint training on multiple datasets, *e.g.*, general and transparent object segmentation datasets for our assistive system, or normal and adverse driving scene segmentation datasets for automated vehicles. The description of dual-head decoder in the assistive system will be detailed in Sec. III-C. The feature maps representation in two decoder heads are illustrated as $\{F_1^1, F_2^1, F_3^1, F_4^1\}$ and $\{F_1^2, F_2^2, F_3^2, F_4^2\}$ in Fig. 2(a). Benefiting from TPM, the amount of GFLOPs and parameters of this dual-head structure is largely reduced compared to deploying two separate models. Mounting multiple decoder heads, it robustifies the feature learned via the shared encoder and prevents overfitting. Meanwhile, the entire model is computationally efficient.

B. Portable Assistive System

The wearable navigation assistance system is composed of a pair of smart vision glasses and a mobile GPU processor, *e.g.*, a lightweight laptop or an NVIDIA AGX Xavier (Fig. 1).

As shown in Fig. 3, the smart vision glasses have been integrated with a RealSense R200 [52] RGB-D sensor to enable real-time acquisition of RGB and depth images at the resolution of 640×480, and a pair of bone-conduction earphones for delivering acoustic feedback to people with

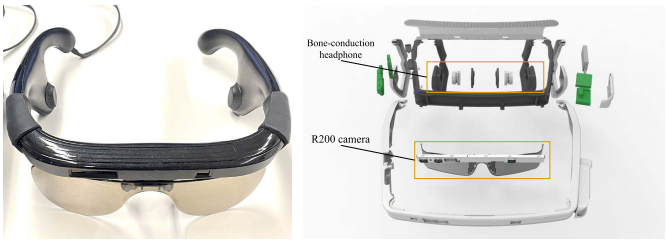


Fig. 3. Detailed illustration of components in the smart vision glasses designed for assisting visually impaired people. The main components: RealSense R200 camera and bone-conduction headphones.

visual impairments. This is crucial as visually impaired people often rely on the sounds from the surroundings for determining the orientation and bone-conduction headphones will not block their ears when using the assistive system. In texture-less indoor scenes, the projected infrared speckles (Fig. 3) will augment the environments, which are beneficial for stereo matching algorithms (*e.g.*, R200 leverages a straightforward correlation engine [52]) to yield dense depth estimation. In our assistive system, depth information is mainly used to assist the obstacle avoidance function, *e.g.*, to prioritize near-range objects over mid- and long-range objects.

C. System Algorithm

The algorithm and user interface of our assistive system with the dual-head Trans4Trans are described in Algorithm 1.

1) System Setting: The R200 sensor [52] has a frame rate of up to 60. To guarantee a timely data acquisition and cover the depth range from $0.5m$, we preset the frame rate to 60 and the resolution to 320×240. Then, the camera can obtain sufficiently-accurate depth information and cover near-range objects. Once the system starts, it repeats the algorithm every n seconds. According to our experiments, the time interval setting as 1–2 seconds effectively prevents cognitive overload, especially in cases of complex scenes containing a large number of objects. Still, it is adjustable depending on the demands, *e.g.*, a short interval for more feedback to explore unknown space. Within 2 seconds, our efficient Trans4Trans model can perform segmentation of approximately 20 frames.

2) Obstacle Avoidance: When moving indoors with limited space, the building materials and densely arranged objects will seriously hinder the flexibility of using the common white cane as an obstacle avoidance aid. To prevent collisions, we preset the highest priority for obstacle avoidance. In contrast to the fixed and limited categories defined in an obstacle avoidance engine [29], we leave the obstacles as open-set and detect them based on the depth information $D \in \mathcal{R}^{H\times W}$. In other words, if the average value of the depth information $\overline{D}$ is smaller than the preset threshold $\theta_{obstacle}$, the user will be immediately notified via *vibrations*. To minimize the uncertainty of vibrations and preclude the chaotic and low-confidence segmentation from the less-textured images, only one single default threshold is set. According to our pre-tests and the minimum effective range ($0.5m$) of the R200 camera, we set the distance threshold as $\theta_{obstacle} = 1m$, and thereby the system can effectively detect open-set and near-range obstacles and output *vibration* hints.

Algorithm 1 Assistive System

Data: RGB-D as $X \in \mathcal{R}^{H\times W\times 3}$ and $D \in \mathcal{R}^{H\times W}$.
Result: General $G \in \mathcal{R}^{H\times W\times 13}$; Transparency $T \in \mathcal{R}^{H\times W\times 11}$;
Initialize walkable rates: R_l, R_f, R_r, parameters: $\theta_{obstacle}, \theta_{trans}, \theta_{walkable}$;
while *system start and each n seconds* **do**
  RGB-D update and Trans4Trans segmentation:
  $G_{path} \in \mathcal{R}^{H\times W}, G_{object} \in \mathcal{R}^{H\times W\times 12}$;
  $T_{stuff} \in \mathcal{R}^{H\times W\times 3}, T_{thing} \in \mathcal{R}^{H\times W\times 8}$;
  partitions $\{R_l, R_f, R_r\} \leftarrow G_{path}$;
  if $\overline{D} < \theta_{obstacle}$ **then**
    $vibration$ as obstacle warning;
  else if $max\{\overline{T_i}\} \in T_{stuff} > \theta_{trans}$ **then**
    $speech \leftarrow argmax\{\overline{T_i}\} \in T_{stuff}$;
  else if $max\{R_l, R_f, R_r\} > \theta_{walkable}$ **then**
    $speech \leftarrow argmax\{R_l, R_f, R_r\} \in \{left, forward, right\}$;
  else
    $speech \leftarrow nearest\{T_{thing}, G_{object}\}$;
  end
end

3) (Transparent) Object Segmentation: After receiving the RGB image $X \in \mathcal{R}^{H\times W\times 3}$, our dual-head Trans4Trans model outputs two segmentation predictions, which are general object segmentation $G \in \mathcal{R}^{H\times W\times 13}$ and transparent object segmentation $T \in \mathcal{R}^{H\times W\times 11}$, respectively. The categories will be introduced later in the dataset description. The general object segmentation is divided into G_{path} of *walkable path*, *i.e.*, *floor* class, and G_{object} of other *objects* [15]. Besides, the transparent object segmentation is divided into two disjoint sets as: $T_{stuff} \in \mathcal{R}^{H\times W\times 3}$ with {*window, glass door, glass wall*}, and $T_{things} \in \mathcal{R}^{H\times W\times 8}$ with {*shelf, jar/tank, freezer, eyeglass, cup, bowl, bottle, box*} [16]. When combining G_{path} and T_{stuff}, we preset a higher priority of prompts for transparent objects. Specifically, when the maximum segmented area of transparent stuff is greater than a preset threshold θ_{trans}, its corresponding category is fed back in a speech form, before that of other general objects. Based on our experiments, if the whole image area is 1.0, we set the $\theta_{trans} = 0.5$ to eliminate the effects of jitter-errors in segmentation.

4) Walkable Path Detection: After achieving general object segmentation, the walkable mask G_{path} is further partitioned into three regions as $\{left, forward, right\}$ directions for orientation assistance. The local ratio R is calculated by G^i_{path}/A^i_{image}, where i represents one of the three directions and A denotes the image area. As a result, the horizontally divided ratios are denoted as $\{R_l, R_f, R_r\} \leftarrow G_{path}$. Then, an intuitive and effective strategy is to prompt the direction that has the largest walkable area, only when the largest local ratio R is greater than the preset threshold $\theta_{walkable}$. In this case, we set a strict threshold $\theta_{walkable} = 0.4$ to ensure that the maximum area is safe enough and walkable for the user. According to our test, this orientation approach guarantees anti-veering in a straight path outdoors and indoors. Furthermore, it can accurately predict the best instantaneous turning direction during walking at an intersection, so as to constantly yield a safe direction suggestion.

TABLE I

COMPARISON ON STANFORD2D3D [15] AND TRANS10K-V2 [16]. THE DECODER EMBEDDING DIMENSION OF TRANS4TRANS IS 64, AND IS {128, 128, 256} FOR TRANS2SEG-T/-S/-M [16], RESPECTIVELY

Network	Encoder	Decoder	GFLOPs	MParams	Stanford2D3D	Trans10K-v2
Trans2Seg-T [16]	PVT-T [46]	Transformer [16]	10.16	13.11	41.00	64.60
Trans2Seg-S [16]	PVT-S [46]		19.58	24.36	41.89	68.47
Trans2Seg-M [16]	PVT-M [46]		49.00	56.20	42.49	72.10
Trans4Trans-T	PVT-T [46]	TPM / Single-head	10.45	12.71	41.28(+0.28)	68.63(+4.03)
Trans4Trans-S	PVT-S [46]		19.92	23.95	44.47(+3.04)	74.15(+5.68)
Trans4Trans-M	PVT-M [46]		34.38	43.65	45.73(+3.24)	75.14(+3.04)
Trans4Trans-T	PVT-T [46]	TPM / Dual-head	11.22	13.10	40.44(-0.56)	69.84(+5.24)
Trans4Trans-S	PVT-S [46]		20.69	24.34	43.45(+1.56)	74.57(+6.10)
Trans4Trans-M	PVT-M [46]		35.17	44.04	45.15(+2.66)	74.98(+2.88)

IV. EXPERIMENTS

A. Datasets and Settings

1) Trans10K-v2: [16] has 5000, 1000, and 4428 images for training, validation, and testing, respectively. The $H \times W$ resolution of the images is 835×1113. There are 11 categories marked as *shelf, jar or tank, freezer, window, glass door, eyeglass, cup, wall, glass bow, water bottle*, and *storage box*.

2) Stanford2D3D: [15] has 13 categories: *beam, board, bookcase, ceiling, chair, clutter, column, door, floor, sofa, table, wall*, and *window*. The image resolution is 1080×1080. Based on the fold-1 splitting [15], the training set has 52905 images from *Area 1-4* and *Area 6*, the validation set has 6261 images from *Area 5a*, and the test set has 11332 images from *Area 5b*.

3) Cityscapes: [17] is captured under *normal* driving conditions and has 19 categories. It has 2979 and 500 images for training and validation. The image resolution is 2048×1024.

4) ACDC: [18] is captured under four *adverse* driving conditions (*fog, night, rain,snow*) and has 19 categories. There are 1600 training- and 406 validation images which are publicly available, and 2000 test images for benchmarking. The resolution of the images is 1920×1080.

5) DADA-Seg: [19] is a testing dataset captured in traffic *accidental* scenes and has 19 categories. The image resolution is 1584×660. The dataset is utilized together with ACDC to study the robustness of Trans4Trans in adverse conditions.

6) Implementation Details: Our Trans4Trans model is implemented with CUDA 11.2 and PyTorch 1.8.0 with an initial learning rate of $1e^{-4}$ and scheduled by the poly strategy [53] with power 0.9 in 100 epochs. AdamW [54] is the optimizer with epsilon $1e^{-8}$ and weight decay $1e^{-4}$, and batch size is set as 4 on each of four 1080Ti GPUs. The experiments for ablating the effect of embedding channels are conducted on a single GPU. The images in the training and testing stages are resized in the resolution of 512×512 or 768×768 (will be specified) for the experiments, to maintain the shape of position embedding. For a fair comparison with Trans2Seg [16], some tricks such as OHEM, auxiliary and class-weighted losses are not applied in our experiments. GFLOPs are calculated at 512×512 unless otherwise specified.

B. Accuracy of General and Transparent Object Segmentation

1) Results: The comparison of encoder-decoder model architectures is shown in Table I. Compared with Trans2Seg [16], the encoder and decoder in our approach are both transformer-based. The single- and dual-head decoders are constructed by our TPM. As shown in Table I, the single-head Trans4Trans clearly outperforms Trans2Seg in all settings (T, S, and M are short for Tiny, Small, and Medium). Trans4Trans-M achieves the best performance in mIoU with 45.73% on Stanford2D3D dataset and 75.14% on Trans10K-v2 dataset, exceeding by more than 3% w.r.t. Trans2Seg-M on the transparent object segmentation benchmark. Meanwhile, compared to Trans2Seg-M, the computational complexity in GFLOPs is much smaller in our approach Trans4Trans-M, while Trans4Trans-T and Trans4Trans-S also achieve better performance than the corresponding Trans2Seg variant on Trans10K-v2.

When incorporating more general knowledge by learning jointly with supervision from Stanford2D3D dataset, dual-head Trans4Trans consistently improves the performance on Trans10K-v2 compared with Trans2Seg. More importantly, dual-head Trans4Trans highly alleviates the problem of overfitting and reduces false positives of transparent obstacle warning observed in our field tests, and thereby it is more suitable for real-world navigation perception. Overall, the superiority and efficiency of Trans4Trans are verified through the experiments on transparent- and general object segmentation.

2) Combination of CNN / Transformer: In Table II, varied combinations of different encoders and decoders are compared, where FCN [39] and OCNet [41] are based on CNN, whereas Trans2Seg is composed of CNN-based encoder and transformer-based decoder. In contrast, Trans4Trans is a fully transformer-based encoder-decoder model. Our approach

TABLE II

EFFECT OF CNN/TRANSFORMER COMBINATIONS ON TRANS10K-V2

Method	Trans. Enc.	CNN Enc.	Trans. Dec.	CNN Dec.	GFLOPs↓	mIoU (%) ↑
FCN [39]		✓		✓	42.2	62.7
OCNet [41]		✓		✓	43.3	66.3
Trans2Seg [16]		✓	✓		40.9	69.2
PVT [46]	✓		✓		49.0	72.1
Trans4Trans (ours)	✓		✓		**34.3**	**75.1**

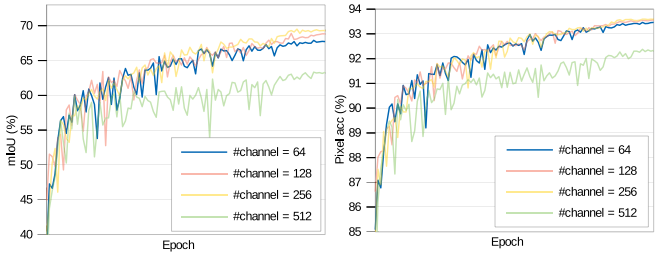


Fig. 4. mIoU and pixel accuracy curves of 4 TPM channel settings.

TABLE III

EFFECT OF TPM CHANNEL. MODELS ARE TRAINED ON TRANS10K-V2 AT 512×512 ON ONE GPU. ACC DENOTES PIXEL ACCURACY

Channel	MParams	GFLOPs	Acc (%)	mIoU (%)
64	10.45	12.71	93.46	67.89
128	12.46	14.02	93.49	68.88
256	20.41	17.82	93.58	69.51
512	51.50	33.82	92.37	63.33

TABLE IV
COMPARISON OF STATE-OF-THE-ART MODELS ON TRANS10K-V2 [16]

Method	GFLOPs ↓	Acc ↑	mIoU ↑	Category IoU ↑											
				Background	Shelf	Jar/Tank	Freezer	Window	Door	Eyeglass	Cup	Wall	Bowl	Bottle	Box
FastSCNN [55]	1.01	88.05	51.93	90.64	32.76	41.12	47.28	47.47	44.64	48.99	67.88	63.80	55.08	58.86	24.65
HRNet_w18 [20]	4.20	89.58	54.25	92.47	27.66	45.08	40.53	45.66	45.00	68.05	73.24	64.86	52.85	62.52	33.02
LEDNet [56]	6.23	86.07	46.40	88.59	28.13	36.72	32.45	43.77	38.55	41.51	64.19	60.05	42.40	53.12	27.29
Trans4Trans-T	10.45	93.23	68.63	94.44	48.39	61.89	61.86	61.14	54.83	73.60	83.03	75.20	74.69	75.26	59.19
ICNet [57]	10.64	78.23	23.39	83.29	2.96	4.91	9.33	19.24	15.35	24.11	44.54	41.49	7.58	27.47	3.80
BiSeNet [53]	19.91	89.13	58.40	90.12	39.54	53.71	50.90	46.95	44.68	64.32	72.86	63.57	61.38	67.88	44.85
Trans4Trans-S	19.92	94.57	74.15	95.60	**57.05**	71.18	**70.21**	63.95	61.25	81.67	87.34	78.52	**77.13**	81.00	64.88
DeepLabv3+ [37]	37.98	92.75	68.87	93.82	51.29	64.65	65.71	55.26	57.19	77.06	81.89	72.64	70.81	77.44	58.63
OCNet [41]	43.31	92.03	66.31	93.12	41.47	63.54	60.05	54.10	51.01	79.57	81.95	69.40	68.44	78.41	54.65
Trans2Seg [16]	49.03	94.14	72.15	95.35	53.43	67.82	64.20	59.64	60.56	**88.52**	86.67	75.99	73.98	82.43	57.17
TransLab [35]	61.31	92.67	69.00	93.90	54.36	64.48	65.14	54.58	57.72	79.85	81.61	72.82	69.63	77.50	56.43
PSPNet [58]	187.03	92.47	68.23	93.62	50.33	64.24	70.19	51.51	55.27	79.27	81.93	71.95	68.91	77.13	54.43
DANet [40]	198.00	92.70	68.81	93.69	47.69	66.05	70.18	53.01	56.15	77.73	82.89	72.24	72.18	77.87	56.06
Trans4Trans-M	34.38	**95.01**	**75.14**	**96.08**	55.81	**71.46**	69.25	**65.16**	**63.96**	83.84	**88.21**	**80.29**	76.33	**83.09**	**68.09**

outperforms both these aforementioned competitive networks such as OCNet and transformer-based encoder-decoder architectures such as PVT [46]. Besides, Trans4Trans keeps clearly smaller GFLOPs while being more accurate, demonstrating its suitability for efficient transparent object segmentation.

3) Effect of TPM Channel: As TPM is one of our critical designs, the ablation study of different numbers of embedding channels applied in the Trans4Trans decoder is conducted, where the effects are illustrated in Table III and Fig. 4. It reveals that the larger channel number, the better performance, until 256. The drop at 512 indicates that the decoder overfits the encoded feature as shown in Fig. 4 and the computation complexity becomes exceedingly large with the increase of channel number for TPM. For the response time-critical wearable system, we adopt the smallest 64 channels when deploying dual-head Trans4Trans due to its highest efficiency and good segmentation performance. Yet, to pursue high accuracy on driving-scene datasets, we set TPM channel of Trans4Trans-T/-S/-M as {64, 128, 256}, respectively.

4) Comparison to State-of-the-Art Models: In Table IV, different accuracy- and efficiency-oriented semantic segmentation approaches are compared according to [35]. The superiority of Trans4Trans is further confirmed through the listed experimental results in Table IV, compared with both CNNs and transformer-based approaches like Trans2Seg [16]. Our Trans4Trans-M model outperforms the previous best method Trans2Seg by 2.99% in mIoU and 0.87% in Acc, while requiring much less GFLOPs. For category-wise accuracy, our Trans4Trans model achieves state-of-the-art performances in IoU on the classes *background*, *jar or tank*, *window*, *door*, *cup*, *wall*, *bottle*, and *box*, indicating the efficacy of transparent object segmentation of Trans4Trans.

C. Segmentation Robustness in Driving Scenes

Apart from segmenting general and transparent objects, we verify our Trans4Trans model on driving scene datasets to show its effectiveness and potential in ITS applications.

1) Ablation of Dual-Head Trans4Trans: Five groups of results are shown in Table V. Our TPM-based Trans4Trans trained at the 512×512 resolution, illustrates better performance compared with PVT on both driving scene datasets.

TABLE V
COMPARISON ON CITYSCAPES [17] AND ACDC [18]. EMBEDDING DIMENSION OF *model*-T/-S/-M DECODER IS {64, 128, 256}

Network	Encoder	Decoder	GFLOPs	MParams	Cityscapes	ACDC
PVT-T	PVT-T [46]		10.30	13.11	58.09	53.65
PVT-S	PVT-S [46]	Transformer [16]	19.77	24.35	59.68	57.13
PVT-M	PVT-M [46]		36.87	51.83	60.38	58.60
Trans4Trans-T	PVT-T [46]		10.45	12.71	60.41(+2.32)	54.37(+0.72)
Trans4Trans-S	PVT-S [46]	TPM / Single-head	21.98	25.00	63.08(+3.40)	60.70(+3,57)
Trans4Trans-M	PVT-M [46]		44.38	48.77	65.63(+5,25)	61.91(+3,31)
Trans4Trans-T	PVT-T [46]		11.23	13.10	57.42(-0.67)	56.36(+2.71)
Trans4Trans-S	PVT-S [46]	TPM / Dual-head	24.82	26.45	62.39(+2.71)	62.14(+5.01)
Trans4Trans-M	PVT-M [46]		55.16	54.28	63.00(+2.62)	63.88(+5.28)
Trans4Trans-T	PVTv2-B1 [47]		9.18	13.53	63.25(+5.16)	59.25(+5.60)
Trans4Trans-S	PVTv2-B2 [47]	TPM / Single-head	19.27	25.62	67.28(+7.60)	64.61(+7.48)
Trans4Trans-M	PVTv2-B3 [47]		41.89	49.55	69.34(+8.96)	65.92(+7.32)
Trans4Trans-T	PVTv2-B1 [47]		10.00	13.93	62.31(+4.22)	61.86(+8.21)
Trans4Trans-S	PVTv2-B2 [47]	TPM / Dual-head	22.17	27.08	65.98(+6.30)	64.83(+7.70)
Trans4Trans-M	PVTv2-B3 [47]		52.77	55.09	69.05(+8.67)	66.65(+8.05)

On Cityscapes, Trans4Trans-M leveraging PVT encoder outperforms PVT-M by 5.25% and Trans4Trans-M leveraging PVTv2 as the encoder surpasses by 8.96% (with an additional 3.71% gain). On ACDC, our Trans4Trans-M leveraging PVT encoder outperforms PVT-M by 5.28% and the one leveraging PVTv2-M as the encoder exceeds by 8.05% (with an additional 2.77% boost) while utilizing TPM/Dual-head in the decoder architecture. Since ACDC is a dataset containing different adverse conditions, these results evidence that TPM/Dual-head has the better robustness under environment changes in driving scene segmentation, as it incorporates more generalized knowledge learned from diverse images in both datasets.

2) Segmentation in Normal Conditions: In Table VI, results of Trans4Trans trained with the input size of 768×768 are compared with more than 15 state-of-the-art methods[1]. All Trans4Trans are constructed with the best single-head Trans4Trans illustrated in Table V. Our Trans4Trans-M approach with PVTv2-B3 as encoder achieves the best performance with an mIoU of 81.54% on Cityscapes. Compared with the state-of-the-art methods such as SETR [43] and PSPNet [58], our Trans4Trans approach shows smaller GFLOPs (94.25) and less parameters (49.55M), which are relevant for fast inference in automated vehicles. TransTrans-T and -S models with lighter encoder architectures also show

[1]For a fair comparison, model weights are obtained by the same framework MMSegmentation: https://github.com/open-mmlab/mmsegmentation

TABLE VI

COMPARISON ON CITYSCAPES. GFLOPS ARE CALCULATED AT 768×768. "MS" MEANS MULTI-SCALE TESTING

Methods	Encoder	GFLOPs ↓	MParams ↓	mIoU (MS) ↑
FastSCNN [55]	Fast-SCNN	2.07	1.46	72.65
CGNet [59]	CGNet-M3N21	7.72	0.50	64.80
Trans4Trans-T	PVTv2-B1 [47]	20.66	13.53	78.23
HRNet [20]	HRNetV2p-W18s	21.70	3.94	77.48
SegFormer-B1 [45]	MiT-B1	29.85	13.66	78.43
ERFNet [6]	ERFNet	30.22	2.07	72.10
Trans4Trans-S	PVTv2-B2 [47]	43.37	25.62	80.02
PSPNet [58]	MobileNetV2	119.09	13.72	70.20
PSPNet [58]	ResNet-18	119.27	12.77	76.90
SegFormer-B2 [45]	MiT-B2	127.86	27.33	80.46
SegFormer-B3 [45]	MiT-B3	160.78	47.18	81.50
DeepLabv3+ [37]	MobileNetv2	169.53	18.70	75.20
HRNet [20]	HRNetV2p-W48	210.57	65.86	80.72
PSPNet [58]	ResNet-50	401.51	48.98	79.96
PSPNet [60]	ResNet-101	573.48	67.95	80.04
SETR-Naive [43]	ViT-L [42]	698.52	306.58	77.90
SETR-MLA [43]	ViT-L [42]	712.76	310.81	77.24
SETR-PUP [43]	ViT-L [42]	818.26	319.11	79.34
Trans4Trans-M	PVTv2-B3 [47]	94.25	49.55	**81.54**

TABLE VII

COMPARISON ON ALL-ACDC [18] AND ALL-DADA [19] CONDITIONS. CS: CITYSCAPES [17]. GFLOPS@768×768

Method	Trained on	GFLOPs ↓	Fog	Night	Rain	Snow	All-ACDC	All-DADA
DeepLabv3+ [37]	CS	178.1	45.7	25.0	50.0	42.0	41.6	10.4
HRNet [20]	CS	210.5	38.4	20.6	44.8	35.1	35.3	15.5
Trans4Trans-M	CS	41.8	**74.1**	**31.1**	**63.4**	**57.9**	**55.7**	**27.7**
DeepLabv3+ [37]	ACDC	178.1	69.1	60.9	74.1	69.6	70.5	26.8
HRNet [20]	ACDC	210.5	74.7	**65.3**	**77.7**	76.3	75.0	27.5
Trans4Trans-M	ACDC	41.8	**79.8**	55.3	77.4	**78.6**	**75.2**	**32.4**
Trans4Trans-M	ACDC+CS	41.8	**81.4**	56.0	77.0	**78.8**	**76.3**	**39.2**

high scores of 78.23% and 80.02% in mIoU. The lightest Trans4Trans outperforms state-of-the-art efficient networks FastSCNN [55] and CGNet [59] by large margins, and it achieves a similar score as SegFormer [45] while being significantly more efficient.

3) Segmentation in Adverse Conditions: In Table VII, we test Trans4Trans-M on ACDC [18] and DADA-seg [19], which are composed of adverse- and accidental scenes, respectively. The results of Trans4Trans are obtained via MMSegmentation with a resolution of 768 × 768. In the first group, Trans4Trans-M obtains higher performances of 55.7% and 27.7% in mIoU when compared with HRNet [20] and DeepLabV3+ [37]. Trans4Trans outperforms them in all four adverse conditions and accidental scenes, which demonstrates its high generalization capacity to these unseen domains. This is because with both transformer-based encoder and decoder, Trans4Trans can associate long-range visual concepts for robustly inferring semantics, despite local texture- and illumination changes in different scenarios like nighttime and accident scenes. Thanks to our efficient backbone, Trans4Trans surpasses HRNet by >20% and >12% on All-ACDC and All-DADA with only its 20% GFLOPs. In the second group of Table VII, Trans4Trans again indicates better overall performances on two datasets. Finally, the model trained on ACDC and Cityscapes shows the best overall scores on All-ACDC and All-DADA with 76.3% and 39.2% in mIoU, respectively, illustrating that co-training on normal and adverse data can improve the performance of the model under both adverse and extreme accident conditions.

TABLE VIII

INFERENCE TIME (MS/FRAME) OF DUAL-HEAD TRANS4TRANS IS TESTED IN HALF-/SINGLE-PRECISION ON VARIOUS GPUS AT 512×512

Network	NVIDIA Xavier (ms) ↓	MX350 (ms) ↓	RTX 2070 (ms) ↓
Trans4Trans-M	115.9(±1.1) / 202.8(±1.1)	186.1(±0.3) / 243.2(±0.3)	22.9(±0.3) / 36.6(±0.8)
Trans4Trans-S	95.3(±0.6) / 158.6(±1.8)	140.6(±0.3) / 188.4(±0.4)	17.1(±0.3) / 27.7(±0.5)
Trans4Trans-T	75.8(±0.7) / 122.7(±0.7)	101.5(±0.3) / 141.7(±1.6)	12.8(±0.5) / 20.3(±0.5)

D. Real-Time Performance

To measure the efficiency of Trans4Trans, 300 samples from Trans10K-v2 with a batch size of 1 and a resolution of 512×512 are tested on three GPUs, *i.e.*, an NVIDIA AGX Xavier in the MAXN mode, an NVIDIA GeForce MX350 on a lightweight laptop, and an RTX 2070 on a workstation. As shown in Table VIII, the running time (latency) of Trans4Trans-T are considerably lower than the other two versions, meanwhile the performances of three models on both datasets are suitable for our system. In real applications, the more timely response of the navigation system is beneficial for assisting users with a similar prediction accuracy on each frame. Hence, the tiny version model is selected and deployed on the lightweight laptop (with an MX350 GPU) to conduct the user study. The entire system takes 0.14 ms ($\pm$0.11) for the image acquisition, 101.5 ms ($\pm$0.3) for segmentation, and 74.47 ms ($\pm$2.41) for the assistive algorithm. Thus, the entire system requires a total of 176ms.

E. Transparent Feature

1) Analysis of Contexts, Reflections, and Backgrounds: To investigate the impact of context information (such as door frames and walls) or reflection features on transparency perception, we visualize the segmentation results from six scales (from 100% to 20%) of the original images in Fig. 5. In Fig. 5(a), among all six scales, even in the 20% case with less context, two overlapping doors are accurately segmented. In Fig. 5(b), the sneeze guard is recognized as the glass wall, since it has the glass-like reflection. Another reason is the sneeze guard has no frame and overlaps with the background wall. Therefore, the background of transparent objects affects the object classification if they lack contextual information. In the 40% ratio in Fig. 5(c), with only a one-side outer frame and part of the reflection, it can still segment the area of the glass wall. However, in the 20% ratio, it is confused due to the tiny frame and absence of any reflections. The errors in the latter two ratios of Fig. 5(d) are caused by the lack of texture information and reflections. Based on the analysis of visualizations, three insights are provided: (1) The contextual information, *e.g.*, the outer frame, is a vital factor for the transparency segmentation; (2) The reflection characteristic of glass or transparent objects is crucial; (3) The background texture of transparent objects also interferes with the segmentation results when they lack contextual information. Thanks to the symmetrical encoder-decoder structure,

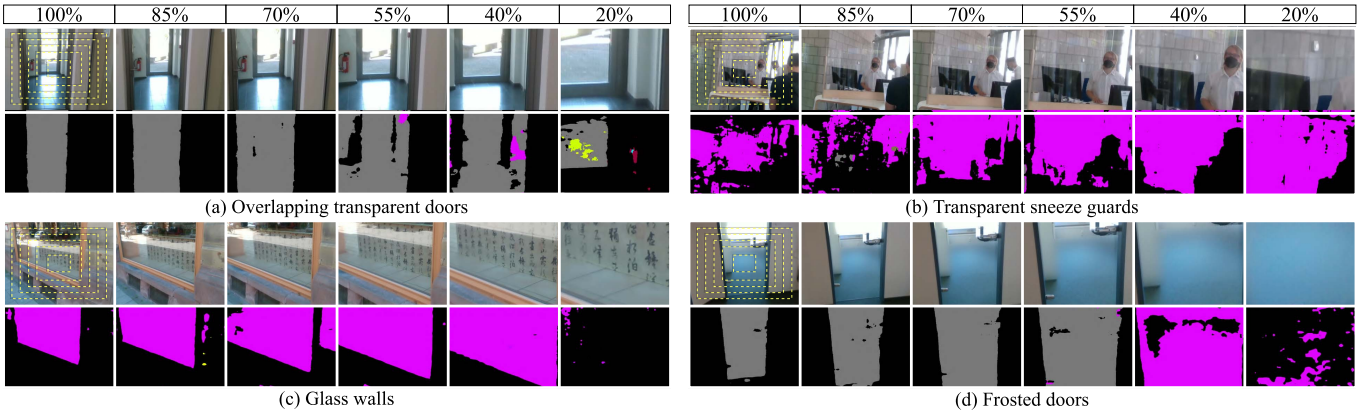


Fig. 5. Visualization of segmentation results from different cropped regions based on image center. Images of six scales from 100% to 20% are cropped from its original images and are separately segmented, in order to ablate the effect of image context.

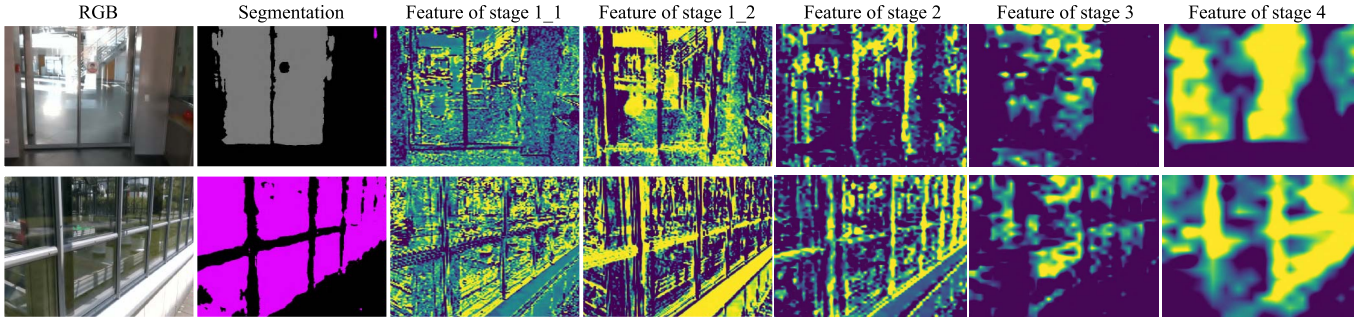


Fig. 6. Visualization of feature maps, corresponding to four different stages in the TPM decoder. The two feature maps from the same stage (stage 1_1 and stage 1_2) indicate the global and local activated features of transparent objects.

Trans4Trans can robustly segment transparent objects even with diminishing context cues in most of the complex real-life scenes.

2) Features Parsing: To investigate the feature parsing capability of the TPM decoder, we visualize the interpreted feature maps of its four stages, as shown in Fig. 6 with real-world scenes. We get two observations: (1) In various low- and high-level stages, the parsed features contain both fine-grained and contextual information, thanks to our Trans4Trans architectural design which enables to capture long-range context priors from very first layers. For example, the fine feature in the stage 1_2 of the first row contains the instance-level information such as *glass door*. The contextual feature in stage 4 contains the precise boundary of the *glass door* as well. Both types of features are critical for the segmentation task; (2) In a same stage, the fine-grained feature is simultaneously reflected in the object and its surrounding. For example, the feature in stage 1_1 of the first row is more inclined to the surrounding, and the one in stage 1_2 is to the area of the *glass door*, and vice versa in the second row.

F. Qualitative Segmentation Analysis

1) Visualization of Transparency Segmentation: Fig. 7 shows qualitative comparisons between Trans4Trans-T and the previous best approach Trans2Seg [16]. In spite of high accuracy scores they have, Fig. 7(a) illustrates some failure cases of both models. Fig. 7(b) shows examples where our model predicts the correct label, but Trans2Seg is confused, indicating the reliability of our approach. In Fig. 7(c)(d), Trans4Trans is not only effective in detecting navigation-critical *glass door* and *glass window*, but can also predict more refined segmentation of small objects like *jar/tank* and *glass cup*.

2) Visualization of Driving Scene Segmentation: In Fig. 8, we visualize the predictions of Trans4Trans* trained on Cityscapes and ACDC, in comparison to DeepLabv3+ [37], HRNet [20], and our Trans4Trans models only trained on ACDC. DeepLabv3+ and HRNet produce noisy results in complex conditions, like the *cars* in shadow (the first row). In adverse weather and week illumination conditions, previous methods yield less precise and even fragmented semantics, like the *trucks* in foggy and rainy scenes (the second and fourth rows) and the *sidewalks* in night and snowy scenes (the third and fifth rows). In accident scenes, which are safety-critical for automated vehicles, existing models cannot generate reliable predictions to be propagated to upper-level applications, as the close *pedestrian* is even completely recognized as *road*. In contrast, Trans4Trans, which learns to gather long-range dependency from the very first layers, delivers more robust segmentation in various scenes, as it is less affected by local texture and illumination changes. Trans4Trans trained on both adverse and normal datasets further improves the performance, resulting sharp and fine-grained semantic segmentation.

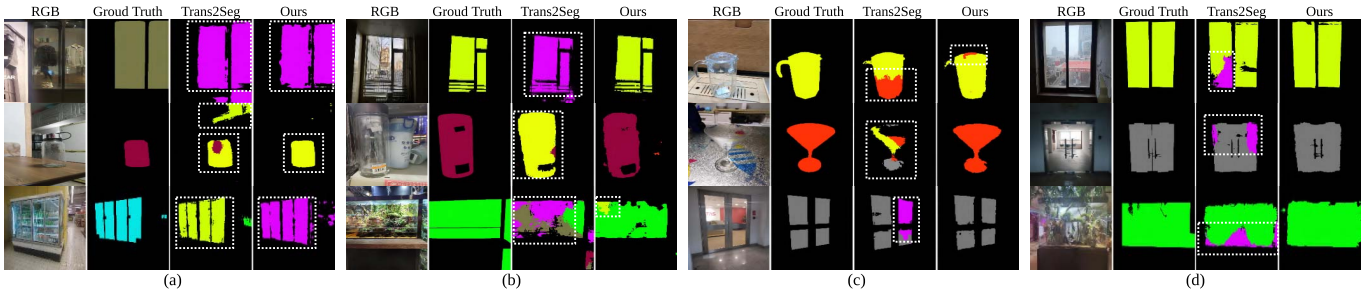


Fig. 7. Qualitative analysis on Trans10K-v2 test set. (a) shows some negative predictions from both models. In (b), our Trans4Trans can correctly segment those cases failed by Trans2Seg. In (c) and (d), our results are more precise.

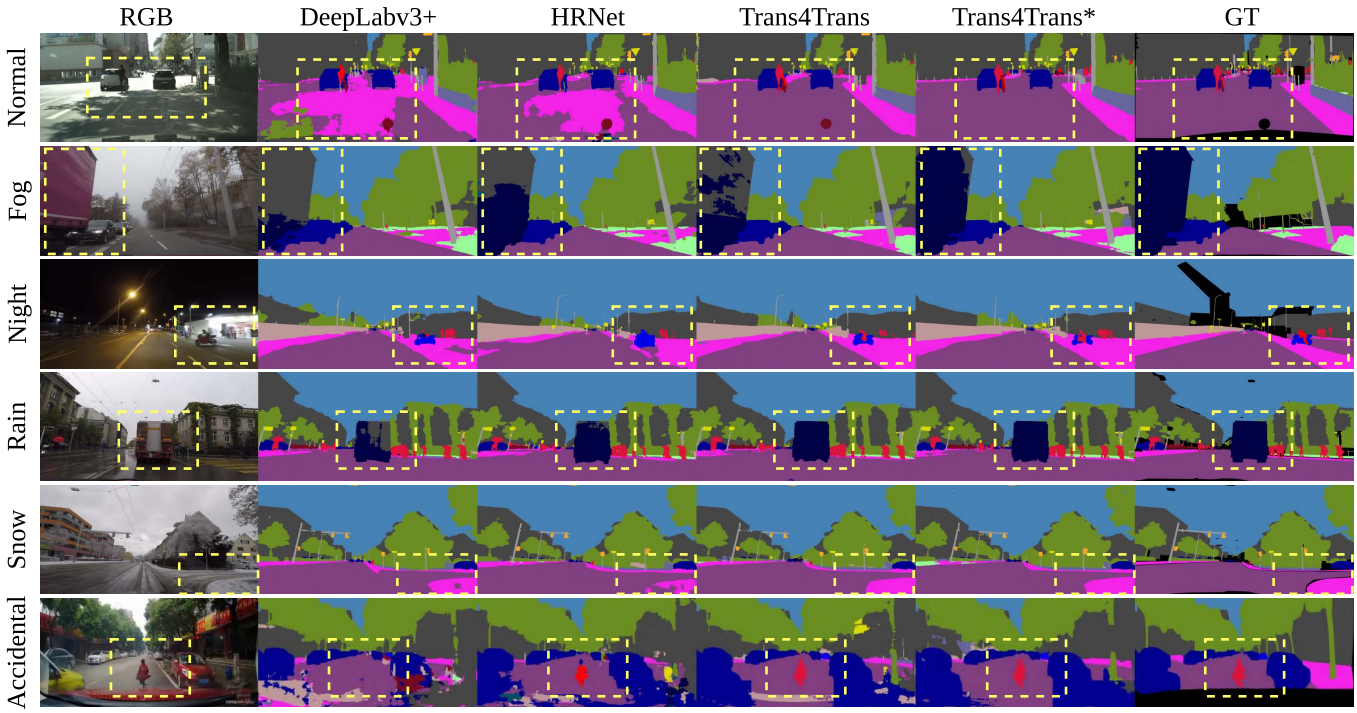


Fig. 8. Qualitative analysis on Cityscapes [17] (*Normal*), ACDC [18] (*Fog, Night, Rain*, and *Snow*), and DADA-seg [19] (*Accidental*). The Trans4Trans* is trained on ACDC+Cityscapes, whereas other models are trained on ACDC dataset.

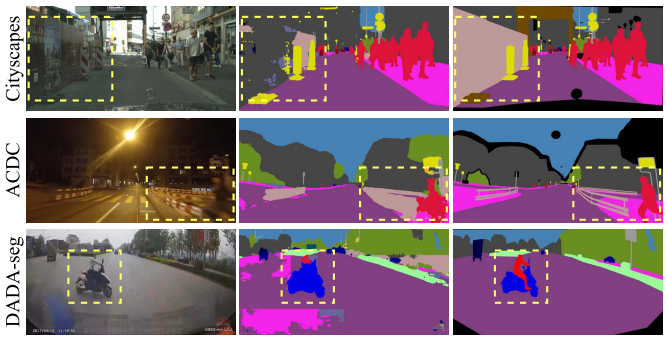


Fig. 9. Failure analysis of driving scene segmentation. From left to right are RGB images, segmentation results, and the ground truth.

3) Failure Analysis of Driving Scene Segmentation: In Fig. 9, some erroneous segmentation cases of the Trans4Trans* model on all three driving datasets are presented. Although the *pedestrian* in the first row can be segmented accurately, the segmentation of the *fence* is less complete. In the second and third rows, the adverse and accidental scenes including extreme motion blur, very weak illuminations, and abnormal behaviors are still challenging for the model. To tackle these issues, fusing such as depth and thermal sensors or other data modalities is a potential solution for further research.

V. User Study

We conducted a qualitative study in order to "understand people's needs and the context within which our future technology might be used" [61]. Another goal was to draw design conclusions for future works. Because system-level performance and efficiency have been evaluated in Sec. IV-D, we did not evaluate any additional objective metrics, but focus on user comments and suggestions on the system. Since the target user group is a very heterogeneous one, and participants with visual impairments are difficult to recruit, we decided

Fig. 10. Incidences of participants using the system for navigation outdoors and indoors.

to evaluate with experts in accessibility and assistive systems with focus on visual impairment.

A. Methodology

As mentioned in Algorithm 1 and Sec. III-C, three main functions are delivered by the assistive system. The system's battery life was approximately 4 hours. Participants tried the system inside 2 buildings (one building is mostly glazed), and the blind expert (E1B) also walked with the system on a 700m route outdoors - see Fig. 10. The study lasted between 30-45 minutes (E5-E8) and 90 minutes (E1B). After a short introduction, the participants put on the system and walked around the rooms, thinking out loud [62]. At the end, demographics and NASA Raw Task Load Index (RTLX) [63] questionnaires were filled in.

B. Participants

In a first step, we evaluated with 5 participants [21], one of whom was an expert, and another one was expert and visually impaired user at the same time. We subsequently repeated the experiment with 3 further sighted experts, and we only report here the aggregated results from the 5 experts: E1B (early blind expert), E5-E8 (sighted experts). When asked if they can see glass objects, E1B said he can sometimes see some light-dark contrasts, which allows him to perceive closed windows. Windows that open inside the room, however, are very dangerous, according to E1B, as one can get serious head injuries. All sighted participants said they can see glass objects, but some of them, like glass doors, glass walls or windows, can be challenging under particular conditions (E5).

C. Cognitive Load

The RTLX, averaged over the five expert participants, was 16.3 with a standard deviation of 8.1. The range is from 0 to 100, the lower the better. This score is enough to keep the user motivated, while not burdening too much [64]. This score, however, must be critically interpreted, since it might not be representative for the users wearing the system in their daily activities. Instead, this score might reflect the cognitive load of the experts assessing the system, since this was their task, and not simulating user behavior. Only the score of the visually impaired participant is highly relevant for the cognitive load of users wearing the system. This score is 13.3, thus very close to the average, but being alone, it has hardly any statistical relevance. More studies will have to be performed in the future to assess the cognitive load of the users wearing the system.

D. User Comments

A thematic analysis [65] performed on the comments made by the experts (both recorded and from the questionnaires) yielded the following insights:

1) Functionality: All experts found the system useful and were impressed by its functionality, for instance, *"for the first time, I had the feeling that artificial intelligence can be useful […]. [I liked] how much it recognized correctly. […] Systems react much better [than 10 years ago…]. I think it's just cool!"* (E1B); Most positive comments are on the amount and type of objects recognized (E1B, E5, E7, E8). The experts gave some important suggestions on subsequent system development, such as identifying more objects (E1B, E5), mounting a second camera to detect low-lying obstacles (E6), and hinting the directions of detected objects (E1B, E5-E7). Two experts (E5, E7) commented positively upon the free path detection, and mentioned that the obstacle detection should be improved. Most issues with the obstacle detection came from the 2$seconds$ cycles (frame aggregation), which often caused a delay and delay inconsistencies (E1B, E5, E6, E7). To tackle this problem, it is desirable to further decrease the system response time. Regarding the suggestions from E5 and E8, adaptive feedback cycles for different functions can be implemented. For example, the feedback of obstacle detection should be given generally faster than for the other two functions. The default in this case could be for instance 1$second$ instead of 2. Besides, the distances of detected objects are helpful for keeping social distances in COVID-19 pandemic times (E6, E8). Three experts (E5-E7) considered that this system is a nice complement to the white cane, but should not be used as alternative.

2) Hardware: The hardware was perceived as quite light weight (E5, E6), and in any case much better than previous prototypes (E1B) tried out by the experts in the past (at least three out of five had tried similar prototypes in the past). However, two experts (E1B, E5) considered the hardware still too big for a real-world deployment: *"[use] a belt instead of a backpack [and] Bluetooth instead of cables for the glasses"* (E1B), *"ideally, it should run on a phone"* (E5). The camera was perceived by E1B as very comfortable to wear, *"although it is thick"*. Experts E5 and E6 commented positively on the system's battery life.

3) Interface: Four out of five experts thought the interface was very intuitive. Only E8 was neutral with respect to this. E6 liked *"the object announcement. The acoustic signal is easy to follow + easy to understand"*. E1B thought that *"the synthetic voice was very helpful, because it differentiates well from background noise"*. E8 thought that the speech output was appropriate for objects recognition, but for the other two functions, some alternatives would be better, in order to diminish the user's cognitive load.

4) Context of Use: Expert E7 thought the system is good for getting an overview of a new room, but not so good for known rooms. He also suggested implementing objects searching and counting. Both E5 and E7 thought the system can be used for social distancing, but referred to two different functions of the system, namely obstacle detection and free path recognition.

Fig. 11. Augmented results with a HoloLens 2 device.

E5 suggested to use the system also for sighted people for warning when walking while looking at the phone.

5) Control: E7 thinks the user should be in full control of the system, like it is the case with the white cane: *"I can't interact with the system (mute) - the white cane does what I want"*. Both E5 and E8 thought it is important to have the option to turn functions on and off, or switch to different modes (E8).

E. Augmented Reality for Partially Sighted People

As transparent obstacles are usually a threat for people with low vision and even hardly distinguishable for sighted people in some confusing situations, our Trans4Trans model are further tested on a HoloLens 2 device by capturing real-world images around our computer vision laboratory. As displayed in Fig. 11, the challenging and omnipresent transparent objects like *glass door*, *transparent wall*, and *glass window* can be reliably and completely segmented, and the colored segmentation masks can be easily overlaid and naturally projected onto the original RGB images shot by the glasses for rendering augmented- or mixed reality. This field test on another glasses device reveals that our segmentation model is robust across cameras and the proposed Trasn4Trans framework can not only assist visually impaired people, but can potentially help partially sighted people.

VI. Conclusion

In this work, we tackle the challenges of transparent object and semantic scene segmentation via Trans4Trans, an efficient transformer architecture with both transformer-based encoder and decoder. At the heart of our assistive system is Trans4Trans, which precisely segments general- and transparent objects with a Transformer Parsing Module (TPM) integrated in the dual-head structure. It achieves state-of-the-art performances on Trans10K-v2 and Stanford2D3D datasets, while being swift and robust to support safety-critical navigation assistance. Considering the synergy between walking- and driving scene perception for improving traffic safety, Trans4Trans is further verified on driving scene segmentation benchmarks including Cityscapes (favorable conditions), ACDC (adverse conditions), and DADA-seg (extreme accident conditions), demonstrating its efficiency and robustness for real-world transportation applications.

The efficient vision transformer is ported in our wearable system with a pair of smart vision glasses designed to help visually impaired people travel and explore surrounding scenes, where transparent objects widely exist in the real life and impede their mobility. Despite a limited number of participants, an extensive set of analyses from a user study and various field tests evidences that the proposed assistive system is reliable and cognitive-load friendly.

References

[1] Z. Cao, X. Xu, B. Hu, and M. Zhou, "Rapid detection of blind roads and crosswalks by using a lightweight semantic segmentation network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6188–6197, Oct. 2021.

[2] A. Mancini, E. Frontoni, and P. Zingaretti, "Mechatronic system to help visually impaired users during walking and running," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 649–660, Feb. 2018.

[3] C. Stahlschmidt, S. von Camen, A. Gavriilidis, and A. Kummert, "Descending step classification using time-of-flight sensor data," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2015, pp. 362–367.

[4] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen, and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1033–1038.

[5] R. Manduchi and S. Kurniawan, "Mobility-related accidents experienced by people with visual impairment," *AER J., Res. Pract. Vis. Impairment Blindnes*, vol. 4, no. 2, pp. 44–54, Feb. 2011.

[6] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[7] K. Yang, X. Hu, L. M. Bergasa, E. Romera, and K. Wang, "PASS: Panoramic annular semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4171–4185, Oct. 2020.

[8] F. M. Butera, "Glass architecture: Is it sustainable," in *Proc. Int. Conf. 'Passive Low Energy Cooling Built Environ.'*, May 2005, pp. 161–168.

[9] M. Maringer, N. Hauck, and A. Mahdavi, "Suitability evaluation of visual indicators on glass walls and doors for visually impaired people," *Appl. Mech. Mater.*, vol. 887, pp. 519–526, Jan. 2019.

[10] H. Mei *et al.*, "Don't hit me! Glass detection in real-world scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3687–3696.

[11] A. Aladren, G. Lopez-Nicolas, L. Puig, and J. J. Guerrero, "Navigation assistance for the visually impaired using RGB-D sensor with range expansion," *IEEE Syst. J.*, vol. 10, no. 3, pp. 922–932, Sep. 2016.

[12] P.-J. Duh, Y.-C. Sung, L.-Y.-F. Chiang, Y.-J. Chang, and K.-W. Chen, "V-eye: A vision-based navigation system for the visually impaired," *IEEE Trans. Multimedia*, vol. 23, pp. 1567–1580, 2021.

[13] M. Saha, A. J. Fiannaca, M. Kneisel, E. Cutrell, and M. R. Morris, "Closing the gap: Designing for the last-few-meters wayfinding problem for people with visual impairments," in *Proc. 21st Int. ACM SIGACCESS Conf. Comput. Accessibility*, Oct. 2019, pp. 222–235.

[14] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 1–11.

[15] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2D-3D-semantic data for indoor scene understanding," 2017, *arXiv:1702.01105*.

[16] E. Xie *et al.*, "Segmenting transparent objects in the wild with transformer," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1–7.

[17] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[18] C. Sakaridis, D. Dai, and L. Van Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10765–10775.

[19] J. Zhang, K. Yang, and R. Stiefelhagen, "ISSAFE: Improving semantic segmentation in accidents by fusing event-based data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 1132–1139.

[20] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[21] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen, "Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1760–1770.

[22] J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, "Smart guiding glasses for visually impaired people in indoor environment," *IEEE Trans. Consum. Electron.*, vol. 63, no. 3, pp. 258–266, Aug. 2017.

[23] O. Miksik *et al.*, "The semantic paintbrush: Interactive 3D mapping and recognition in large outdoor spaces," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 3317–3326.

[24] L. Horne, J. Alvarez, C. McCarthy, M. Salzmann, and N. Barnes, "Semantic labeling for prosthetic vision," *Comput. Vis. Image Understand.*, vol. 149, pp. 113–125, Aug. 2016.

[25] H. Liu, R. Liu, K. Yang, J. Zhang, K. Peng, and R. Stiefelhagen, "HIDA: Towards holistic indoor understanding for the visually impaired via semantic instance segmentation with a wearable solid-state LiDAR sensor," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1780–1790.

[26] N. Long, K. Wang, R. Cheng, W. Hu, and K. Yang, "Unifying obstacle detection, recognition, and fusion based on millimeter wave radar and RGB-depth sensors for the visually impaired," *Rev. Sci. Instrum.*, vol. 90, no. 4, 2019, Art. no. 044102.

[27] I.-H. Hsieh, H.-C. Cheng, H.-H. Ke, H.-C. Chen, and W.-J. Wang, "Outdoor walking guide for the visually-impaired people based on semantic segmentation and depth map," in *Proc. Int. Conf. Pervasive Artif. Intell. (ICPAI)*, Dec. 2020, pp. 144–147.

[28] J. Watson, M. Firman, A. Monszpart, and G. J. Brostow, "Footprints and free space from a single color image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11–20.

[29] Y. Lin, K. Wang, W. Yi, and S. Lian, "Deep learning based wearable assistive system for visually impaired people," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.

[30] Z. Huang, K. Wang, K. Yang, R. Cheng, and J. Bai, "Glass detection and recognition based on the fusion of ultrasonic sensor and RGB-D sensor for the visually impaired," *Proc. SPIE*, vol. 1079, Oct. 2018, Art. no. 107940F.

[31] A. Okazawa, T. Takahata, and T. Harada, "Simultaneous transparent and non-transparent object segmentation with multispectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4977–4984.

[32] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, "Deep polarization cues for transparent object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8602–8611.

[33] J. Lin, Z. He, and R. W. H. Lau, "Rich context aggregation with reflection prior for glass surface detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13415–13424.

[34] K. Xiang, K. Yang, and K. Wang, "Polarization-driven semantic segmentation via efficient attention-bridged fusion," *Opt. Exp.*, vol. 29, no. 4, p. 4802, 2021.

[35] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," in *Proc. ECCV*, 2020, pp. 696–711.

[36] Y. Cao *et al.*, "FakeMix augmentation improves transparent object detection," 2021, *arXiv:2103.13279*.

[37] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 801–818.

[38] H. He *et al.*, "Enhanced boundary learning for glass-like object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15859–15868.

[39] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[40] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[41] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: Object context for semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 8, pp. 2375–2398, Aug. 2021.

[42] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.

[43] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[44] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NeurIPS*, 2021, pp. 1–14.

[46] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[47] W. Wang *et al.*, "PVTv2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, pp. 1–10, 2022.

[48] K. Yang, J. Zhang, S. Reis, X. Hu, and R. Stiefelhagen, "Capturing omni-range context for omnidirectional segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1376–1386.

[49] X. Dong *et al.*, "CSWin transformer: A general vision transformer backbone with cross-shaped windows," 2021, *arXiv:2107.00652*.

[50] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, 2021, pp. 1–3.

[51] 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.

[52] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel® RealSense™ stereoscopic depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1–10.

[53] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. ECCV*, 2018, pp. 325–341.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[55] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," in *Proc. BMVC*, 2019, pp. 1–9.

[56] Y. Wang *et al.*, "LEDNet: A lightweight encoder-decoder network for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1860–1864.

[57] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. ECCV*, 2018, pp. 405–420.

[58] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[59] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.

[60] M. Yin *et al.*, "Disentangled non-local neural networks," in *Proc. ECCV*, 2020, pp. 191–207.

[61] A. Blandford, D. Furniss, and S. Makri, "Qualitative HCI research: Going behind the scenes," *Synth. Lectures Hum.-Centered Informat.*, vol. 9, no. 1, pp. 1–115, Apr. 2016.

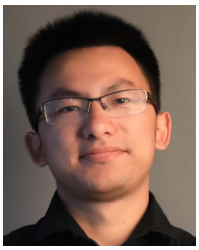
[62] C. J. Johnstone, N. A. Bottsford-Miller, and S. J. Thompson, "Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners," National Center on Educational Outcomes, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. 44, 2006.

[63] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2006, pp. 1–5.

[64] M. Martinez, K. Yang, A. Constantinescu, and R. Stiefelhagen, "Helping the blind to get through COVID-19: Social distancing assistant using real-time semantic segmentation on RGB-D video," *Sensors*, vol. 20, no. 18, p. 5202, Sep. 2020.

[65] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006.

Jiaming Zhang received the B.Sc. degree in computer science and software engineering from Shenzhen University (SZU) in 2015 and the M.Sc. degree in computer science from the Karlsruhe Institute of Technology (KIT) in 2020, where he is currently pursuing the Ph.D. degree. He is a Research Assistant with the Computer Vision for Human–Computer Interaction Laboratory, KIT. His research interests include scene understanding, visual relocalization, and their applications in intelligent vehicles and assistive systems for people with visual impairments.



Kailun Yang received the B.S. degree in measurement technology and instrument from the Beijing Institute of Technology (BIT), the dual degree in economics from Peking University (PKU) in 2014, and the Ph.D. degree in information sensing and instrumentation from the State Key Laboratory of Modern Optical Instrumentation, Zhejiang University (ZJU), in 2019. He performed the Ph.D. degree internship at the RobeSafe Laboratory, University of Alcalá (UAH). He is currently a Post-Doctoral Researcher with the Computer Vision for Human–Computer Interaction (CV:HCI) Laboratoruy, Karlsruhe Institute of Technology (KIT).

Angela Constantinescu received the B.Sc. degree in information technology and the M.Sc. degree in computer science from International University (IU), Germany, in 2006 and 2008, respectively. She is currently a Research Assistant with the Center for Digital Accessibility and Assistive Technology (ACCESS@KIT), Karlsruhe Institute of Technology (KIT). Her research interests include junction between human–computer interaction and accessibility, including audio and tactile user interfaces and evaluation of assistive systems for people with visual impairments.

Kunyu Peng received the B.Sc. degree in automation from the Beijing Institute of Technology (BIT) in 2017, the M.Sc. degree in electrical engineering and information technology from the Karlsruhe Institute of Technology (KIT) in 2021, where she is currently pursuing Ph.D. degree. She is a Research Assistant with the Computer Vision for Human–Computer Interaction Laboratory, KIT. She completed three internships separately at the Chinese Academy of Science, ESME Sudria, and Bosch. Her research interests include human activity recognition, scene understanding, and intelligent vehicles.

Karin Müller received the Diploma (Dipl.-Linguist) and Dr.Phil. degrees from Universität Stuttgart in 1998 and 2002, respectively. She is currently Deputy Director of the Center for Digital Accessibility and Assistive Technology (ACCESS@KIT)—the former Study Center for Visually Impaired Students at Karlsruhe Institute of Technology (KIT). Her research interests include tactile understanding, multimodal interaction, accessibility of documents, and assistive technologies for persons with visual impairments.

Rainer Stiefelhagen (Member, IEEE) received the Diploma (Dipl.-Inform) and Dr.-Ing. degrees from Universität Karlsruhe (TH) in 1996 and 2002, respectively. He is currently a Full Professor of "information technology systems for visually impaired students" with the Karlsruhe Institute of Technology (KIT), where he is the Director of the Computer Vision for Human–Computer Interaction Laboratory, Institute for Anthropomatics and Robotics, and the Center for Digital Accessibility and Assistive Technology. His research interests include computer vision methods for visual perception of humans and their activities, in order to facilitate perceptive multimodal interfaces, humanoid robots, smart environments, multimedia analysis, and assistive technology for persons with visual impairments.