IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 April 2023, accepted 8 June 2023, date of publication 16 June 2023, date of current version 27 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287147

METHODS

Importance Rank-Learning of Objects in Urban Scenes for Assisting Visually Impaired People

YASUHIRO NITTA, (Graduate Student Member, IEEE), MARIKO ISOGAWA, (Member, IEEE), RYO YONETANI, (Member, IEEE), AND MAKI SUGIMOTO, (Member, IEEE)
Graduate School of Science and Technology, Keio University, Yokohama, Kanagawa 223-8522, Japan

Corresponding author: Yasuhiro Nitta (y.nitta@imlab.ics.keio.ac.jp)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT This paper examines an importance rank learning method of objects in urban scenes for assisting visually impaired people. Object detection methods have been used to assist visually impaired people in identifying obstacles in urban scenes, such as cars and trees. However, these existing methods are not dedicated to predicting which obstacle is important. Thus, we propose a method that estimates the importance of objects and warns them to users in order of importance ranking. We introduce a neural network-based ranking estimation method to predict the importance ranking of objects. In particular, our method uses optical flow from the previous frame and region data of detected objects as input. It helps to consider states of moving objects (*e.g.*, cars, motorbikes, people) in a scene. Experimental results show that our model outperforms three other baselines qualitatively and quantitatively. Furthermore, our method was highly evaluated than the baseline methods by qualified caregivers of the visually impaired people.

INDEX TERMS Visually impaired people, object detection, learning-to-rank, differentiable sorting.

I. INTRODUCTION

This paper proposes a method for estimating the importance ranking of the objects in a scene to assist visually impaired/blind (VIB) people. Urban development for VIB individuals has been actively promoted [1], [2]. However, there are still many obstacles (e.g., utility poles, trees, etc.) in the city, requiring VIB people to be assisted by a white cane, a guide dog, or a guide helper when they walk. To overcome this issue, various navigation systems to assist VIB people have been proposed [3], [4], [5], [6]. In particular, advances in computer vision and machine learning techniques have contributed to accurate obstacle detection [7], [8], [9].

Although these existing methods can detect obstacles and landmarks in front of the users, due to the lack of a way to perceive spatial information via vision, VIB people have difficulty understanding the detection results in a short time. Figure 1 shows an example of detection results. This figure indicates obstacles in a city scene detected by an object detection algorithm [10]. While sighted people can perceive the presence of the obstacles as the detection results quickly, VIB people need a method to make the obstacle information accessible in consideration of safety-critical objects in some efficient way other than visual information.

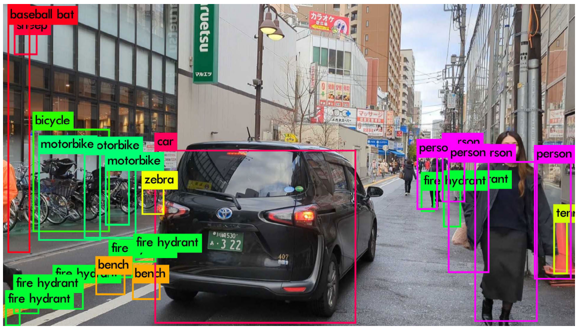


FIGURE 1. Object detection results in a city scene.

Therefore, in this study, we propose a method that estimates the importance ranking of objects in a scene. These inference results help us to transmit necessary information to

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

VIB individuals within a limited time. The order of importance is estimated by a neural network that inputs objects' information, such as the region and class, which are detected from a first-person RGB image and outputs the importance of the object. The proposed framework introduces an importance ranking network with relaxed permutation matrices calculated by NeuralSort [11]. In addition to the object information, the optical flow from the previous frame is added as an input. Optical flow features are used as a clue for obtaining objects' distance and velocity. To evaluate the performance of our method, we constructed a dataset with some sets of an image and the importance ranking of objects in a scene. Using this dataset, we conducted experiments to compare the importance ranking estimated by our method with baselines that determine the ordering based on the area of detected objects, detection confidence, and the distance between the objects and the user. Experimental results showed that the estimated importance order with our method was more appropriate than the other methods. These results show our method's efficacy in warning important objects for VIB users.

To summarize, our contributions are as follows:

- We are the first to propose an importance ranking estimation framework of obstacles for VIB individuals, which contributes to building an effective navigation system for them. This framework is expected to provide a new point of view on walking assistance tasks for VIB individuals.
- We propose a novel framework using optical flow. This framework is intended to estimate the importance ranking of moving objects based on their direction of movement and state. Our simple approach made it possible to understand the priority of warning even if more than two objects are in the same class.
- We describe a neural network-based ranking estimation module to optimize the model toward learning importance order ranking in a scene.
- We provide extensive experiments and show that our method outperforms other baseline methods in ranking estimation and warning audio satisfaction.

II. RELATED WORK

A. WALKING ASSISTANCE FOR VIB PERSON

The important factors for the VIBs in walking are mobility and orientation [12], [13]. Mobility refers to rhythmical walking without stumbling over steps on the ground. Orientation is finding a space that opens in the direction the VIBs should go and move on. VIB individuals use a white cane to acquire mobility and their other remaining senses, such as hearing and haptic perception, to acquire orientation. Guide dogs satisfy both of these perspectives simultaneously. However, the number of guide dogs is limited since training them requires much work. Therefore, Tachi et al. proposed robotic reproductions of a guide dog that can fulfill both aspects [14].

In order to simultaneously acquire mobility and orientation in forms other than guide dogs, walking assistance systems for VIB individuals have been developed. In addition, existing research in recent years has considered using machine learning and computer vision to get highly accurate information. Mahendran et al. [15] proposed a method to estimate traffic conditions and detect moving obstacles with a mobile computing platform. They used smart depth sensors (*i.e.*, OpenCV AI Kit-Depth) to accelerate computations for their tasks. Khan et al. [16] developed an eyeglass-type device with Raspberry Pi equipped with a camera module and ultrasonic sensors. Their proposed device is wearable, and VIB users can use the system with free hands. Sound of Vision [17] proposed a system to assist a VIB person in perception and mobility by presenting three-dimensional information about the surrounding environment through auditory and tactile senses. A tactile device attached to the user's abdomen presented the distance to surrounding objects. Zeng et al. [4] detected obstacles using ambient information collected from a smart white cane. It showed that the path selection with their method outperformed that of a typical white cane in navigation tasks.

These systems above are capable of acquiring ambient information. However, the systems do not consider the priority of warning, such as which detected object should be warned to the user first. We tackle this problem by estimating the importance ranking of objects in a scene. We are the first to focus on the importance ranking of warnings of warning objects on assisting the blind and visually impaired. Our method impacts applications for walking assistance for visually impaired people, especially those aiming at orientation acquisition.

B. IMPORTANCE RANK LEARNING

Importance order estimation for events and environmental information has been considered in the previous works [18], [19]. Such sequential learning has also been used in accessibility research, where Chang et al. proposed a path selection framework for a robot wheelchair with shared autonomy [18]. If the suggested route from this system has high confidence, it selects that route without asking the user for confirmation. On the other hand, if the route is less trusted, the route selection is left to the user. Those kinds of research regarding ranking used Learning-to-Rank (LTR) models.

LTR models learn information rankings according to preference criteria [20]. Common listwise-approach LTR models (*e.g.*, ListNet [21]) map the importance or ranking of each piece of information to a real-valued score and define a loss function that acts directly on this score. These methods need to map like this because it is generally known that the sorting and ranking operations are not differentiable over a large part of the parameter space, making it challenging to learn the rankings in their original form. SoftRank [22] defined scores based on a Gaussian distribution, representing a pseudo-ranking. Yue et al. [23] proposed an LTR method with Support Vector Machine (SVM) to search efficiently for a global optimum. Adarank [24] created weak rankers

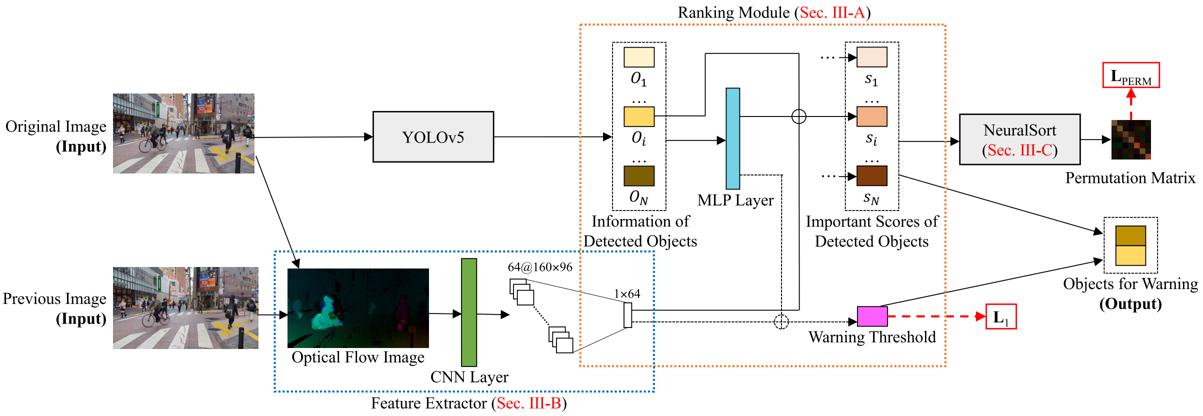


FIGURE 2. **Overview of our ranking estimation method using first-person images. The network extracts time-series features from an optical flow image. Importance ranking scores of objects are estimated with the optical flow features and the information of objects detected from an object detector. This method also estimates the importance score of the warning threshold for each scene. On training, relaxed permutation matrices from NeuralSort were used.**

to improve the performance of ranking prediction. In these methods, Normalized Discounted Cumulative Gain (NDCG) [25], [26], the popular metric to evaluate rankings, is often used to define a loss function. However, this metric is not dedicated to clarifying the relationships of each ranking. RankNet [25] can effectively learn the relationships between pairs of information. LambdaRank [27] extended RankNet by introducing NDCG into the loss function, prioritizing the top objects. However, these two methods are pairwise-approach methods, and there are limitations when considering rankings in the complete list.

Therefore, differentiable sorting methods have been proposed to train the model with the ranking directly on a listwise-approach [11], [28], [29], [30], [31]. These methods can train networks with the original ranking form. Training rankings in their original form make converging to the optimal solution easier. The methods have been used in top-k classification [32] or auction analysis [33]. However, it has not been used in accessibility research. Our method shows that these sorting operators can be effective for the importance ranking estimation of objects in a scene.

III. PROPOSED METHOD

Figure 2 shows our framework to estimate the importance ranking of objects and determine the objects to be warned. This framework outputs the importance ranks and warning thresholds of the objects in the image. First, we input two first-person view images into this framework. One is the current frame image, and the other is the previous one. In addition, we introduce the NeuralSort module that operates a relaxed permutation matrix to train this model using rankings directly.

A. RANKING MODULE

The ranking module outputs an importance score for each object and a threshold of the importance score for warning in the scene, given the detected objects information and optical flow features. Sec. III-B explains optical flow features in detail. Specifically, we input the following object information into the ranking module:

- The coordinates and the area of the bounding box
- The class of the object
- Confidence score based on the object detection model

The coordinates and the area of the objects' bounding boxes help to consider their distances from the user. The importance of objects closer to the user will be higher than those further away. Moreover, the class of object affects the importance score. For example, objects such as bicycles and cars would be more important than people. From another point of view, since we should not present erroneous detection results to the user, we added the detection confidence score to the input.

We estimate the importance ranking score of the *i*-th object (s_i) by considering the information of all detected objects in the scene. The same object information may have different importance depending on what other objects are in the scene. For example, the importance of an object in front of the user may differ if either the left or right side in front of the user is open or both are crowded. Therefore, before estimating the importance score of each object, we obtained the distribution information of objects in the scene from the information of all objects. The MLP layer receives all detection information as input and outputs 100-dimensional object distribution information.

B. FEATURE EXTRACTOR

We used the optical flow image as an input to estimate the importance depending on the state of the moving objects. First, a dense optical flow method (*e.g.*, Farneback method [34]) is performed to generate the optical flow image. Since the flow is calculated for each pixel, the method allows us to obtain the relative velocity of the dynamic objects to the static things such as walls and the ground. The pixel-wise optical flow helps to recognize objects' importance in urban scenes. For example, it is assumed that moving objects (*e.g.*, cars, bicycles) approaching users are more dangerous than those going away. On the other hand, the parked car is less important than that on the road, even if the directions of those cars are the same. In contrast, sparse optical flow methods, such as the Lucas-Kanade algorithm [35], compute flow about detected corners and do not account for relative velocity to static objects. Then a CNN layer aggregates the optical flow image into a feature with 64 dimensions and inputs it to the ranking module.

C. NeuralSort MODULE

We propose to introduce the NeuralSort module to represent the ranking by relaxed permutation matrices. By using these matrices, we can train models by rankings directly. The calculation of relaxed permutation matrices is described in Appendix A. As mentioned in Sec. II-B, listwise ranking learning methods such as ListNet [21] and SoftRank [22] assign arbitrary ranking scores to each rank. It causes the relationship between ranks to be unclear. In this research, we assume that we know the order of importance for each scene, but we do not have prior information about how much ranking score each object has. Therefore, the ranks should be treated directly using a differentiable sorting method such as NeuralSort.

Although it is assumed to learn rankings with unique indices in NeuralSort, our method requires learning rankings with duplicate indices. For practical use, there is an environment in which several objects are equally important. Because of that, training models without considering outputs with the same rankings is not appropriate. We train models with duplicated rankings according to the properties of Eq. 11. If the i-th and j-th objects' rankings are the same, the values in those rows of the matrix are divided equally. For example, if the importance ranking score vector $\mathbf{s} = [7, 3, 3]^T$, then $P_{sort(\mathbf{s})}$ is given by:

$$P_{sort(\mathbf{s})} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \end{bmatrix}, \tag{1}$$

since the ranking scores at the 2nd and the 3rd index are the same.

1) LOSS FUNCTION

The model was optimized with the following two loss functions:

1. **Permutation Loss ($\mathbf{L_{PERM}}$)** : The loss function to train rankings given by:

$$\mathbf{L_{PERM}} = -\sum_{t \in T} \widehat{P}_{z^t} \log P_{z^t}, \tag{2}$$

where $\widehat{P}_{z^t}$ and P_{z^t} are the ground-truth permutation matrix and the estimated permutation matrix, respectively.

2. **L1 Loss ($\mathbf{L_1}$)** : The loss function to optimize the threshold for determining the warning targets. We trained the model to output the importance score threshold of which objects to be warned. The threshold is estimated for each scene. This loss is given by:

$$\mathbf{L_1} = |\hat{v} - v|_1, \tag{3}$$

where $\hat{v}$ and v are the ground-truth threshold value and the estimated threshold value, respectively. As mentioned above, we only have the importance ranking of each object as ground-truth data, and the specific importance scores are unknown in the training phase. Therefore, we optimized the model using the minimum ranking scores of the ground-truth warning objects for each iteration. The warning threshold is expected to be updated according to the changing of objects' ranking scores for each scene.

IV. IMPLEMENTATION DETAILS

As the object detector, we use the pre-trained YOLOv5 [36], known as one of the state-of-the-art methods. In addition, YOLOv5 is structured as a Feature Pyramid Network [37], which extracts features at different scales and resolutions, enabling the network to detect several sizes of objects in the input image. For each detected bounding box, YOLOv5 indicates a detection confidence score from 0 to 1. The object information is highly reliable if the score is close to 1. In this paper, we defined detected objects as those with a detection confidence score of 0.25 or higher.

There exists a dataset for obstacle detection for VIB individuals [38]. However, due to the lack of annotation of the objects (*e.g.*, braille blocks, pedestrian crosswalks) that might be needed for VIB assistance in urban scenes, we prepared an original dataset. Specifically, we trained YOLOv5 by annotating the following classes (person, bicycle, car, motorbike, bus, truck, braille_block, guardrail, crosswalk, signal_red, signal_blue, stairs, tree, bollard, pole, signboard, safety_cone, escalator, grass).

The Adam [39] optimizer was employed at the learning rate of $1e{-}3$. The training typically converges after 200 iterations. While computing the relaxation permutation matrices, we use the temperature parameter $\tau = 1$.

V. EXPERIMENTS

We conducted two experiments to evaluate the performance of our method in importance ranking estimation. First, we evaluated the accuracy of the ranking estimation through a quantitative experiment. Second, we conducted a qualitative experiment to evaluate the effectiveness of assisting VIB users.

A. DATASET

To conduct the experiments, we created an original dataset which is consisted of 272 sets of first-person video frames and the order of importance of the objects in the frames. The objects in each frame were detected by the pre-trained YOLOv5, which are mentioned in Sec. IV. The importance rankings were annotated on frames in 12 videos from Ego4D [40] and five videos from YouTube. The URLs of YouTube videos used for this dataset are described in Appendix B. Those videos are first-person videos showing urban scenes. Table 3 shows the number of frames used. As shown in

TABLE 1. First-person videos used to build the importance ranking dataset.

Video	NumVideos	NumFrames
Ego4D [40]	12	165
YouTube	5	107

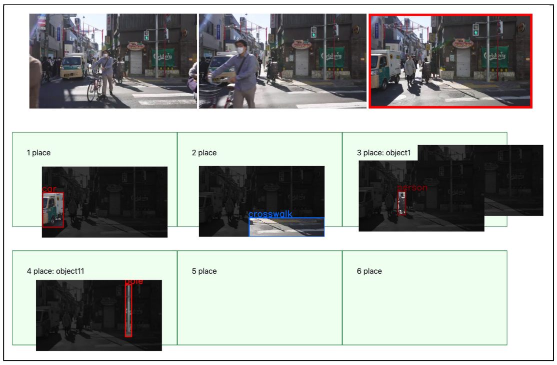


FIGURE 3. Web-based tool to annotate Importance Ranking.

Figure 3, we implemented Web-based annotation tool with FastAPI[1] and JQuery.[2]

During the annotation process, we asked the annotators to rank objects with images highlighted with colored borders (see Figure 3). We instructed the annotators to rank objects that could be important to VIB individuals, assuming they visited the scene in the image for the first time. Detected objects in target scenes were listed by the detection confidence scores of YOLOv5. The listed objects were draggable and ranked by dropping them into a pre-defined box that indicated the rank. In cases there were more than two objects the annotators wanted to rank as the same, the objects were ranked in duplicate by placing multiple highlighted images in the same rank box. The annotators ranked only the important objects and omitted unnecessary ones. The warning threshold was optimized to be equal to the importance ranking score of the lowest-ranked object by the annotators.

The objects were ranked by annotators with qualifications related to the care of the visually impaired. Four annotators (two men and two women) ranked objects for 75 images each, resulting in a collection of 300 annotated data in total. With their informed consent, the annotators were recruited on CrowdWorks,[3] a Japanese cloud-sourcing website.

B. QUANTITATIVE EXPERIMENT

We quantitatively evaluated the accuracy of the importance ranking estimation. We evaluated three baselines and our methods with four metrics.

1) BASELINE METHODS

1. **area** : A method of sorting objects in descending order of the bounding boxes' area. The area of the bounding boxes indicates the distance to the user and the size of the objects. For example, if a car and a person are equally distant from the user, the car would likely be higher ranking to the difference in their cubic volume.
2. **confidence** : A method of sorting objects in descending order of detection confidence scores. Detection confidence scores generally depend on the distribution of objects in the object detection training data. In particular, the detection confidence scores for objects such as people tend to be higher than those for other objects if the images were taken while walking in the city.
3. **closeness** : A method of sorting objects in order of proximity to the user and the object. We defined the proximity by the distance from the lower center of the image to the center of the bounding box.

2) EVALUATION METRICS

We use the following metrics:

1. **Top-N Accuracy** : A metric that measures the probability of the inference result of the 1st-rank object being included up to the N-th rank. We use it to evaluate whether the most important object in the scene was not estimated as low rank. We calculated this metric for the cases N = 1 and 2. This metric is defined as:

$$\mathbf{A}_t = \begin{cases} 1 & \text{if } r_t^j \leq \hat{r}_t^j \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$$\mathbf{A} = \frac{1}{T}\sum_{t \in T} \mathbf{A}_t, \tag{5}$$

where r_t^j and $\hat{r}_t^j$ are the estimated ranking of the j-th object and annotated ranking of the object, respectively.

2. **Top-N Accuracy (Completely)** : A metric that measures the probability that the rank of objects from 1st to Nth is precisely equal to ground-truth order. We use it for evaluating the accuracy of the entire permutation. We calculated this metric for the cases N = 1 and 2. This metric is defined as:

$$\mathbf{ACOMP}_t^j = \begin{cases} 1 & \text{if } r_t^j = \hat{r}_t^j \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

$$\mathbf{ACOMP} = \frac{1}{T}\sum_{t \in T} \prod_{1 \leq j \leq N} \mathbf{ACOMP}_t^j. \tag{7}$$

3. **Normalized Discounted Cumulative Gain (NDCG@5)** : A ranking metric proposed by Burges et al. [25] that measures the proximity of the estimation rankings to the ideal ones. It is the normalized version of the discounted cumulative gain (**DCG@5**). Specifically, This metric is defined as:

$$\mathbf{DCG@5} = \sum_{i=1}^{5} \frac{2^{r^i} - 1}{\log_2(1+i)} \tag{8}$$

$$\mathbf{NDCG@5} = \frac{\mathbf{DCG@5_{pred}}}{\mathbf{DCG@5_{true}}}, \tag{9}$$

[1] Fastapi, https://github.com/tiangolo/fastapi, 20 Nov 2022
[2] Jquery, https://jquery.com/, 20 Nov 2022
[3] CrowdWorks, https://crowdworks.jp/, 20 Nov 2022

TABLE 2. Quantitative results of importance ranking estimation.

Method	Top-N Accuracy		Top-N Accuracy (Completely)		NDCG@5
	$N = 1$	$N = 2$	$N = 1$	$N = 2$	
area	0.432 ± 0.142	0.669 ± 0.171	0.421 ± 0.104	0.153 ± 0.069	0.804 ± 0.074
confidence	0.250 ± 0.070	0.456 ± 0.092	0.228 ± 0.080	0.078 ± 0.047	0.777 ± 0.034
closeness	0.198 ± 0.061	0.286 ± 0.067	0.187 ± 0.064	0.041 ± 0.009	0.685 ± 0.044
Ours	$\mathbf{0.434 \pm 0.127}$	$\mathbf{0.675 \pm 0.170}$	$\mathbf{0.432 \pm 0.095}$	$\mathbf{0.184 \pm 0.079}$	$\mathbf{0.892 \pm 0.027}$

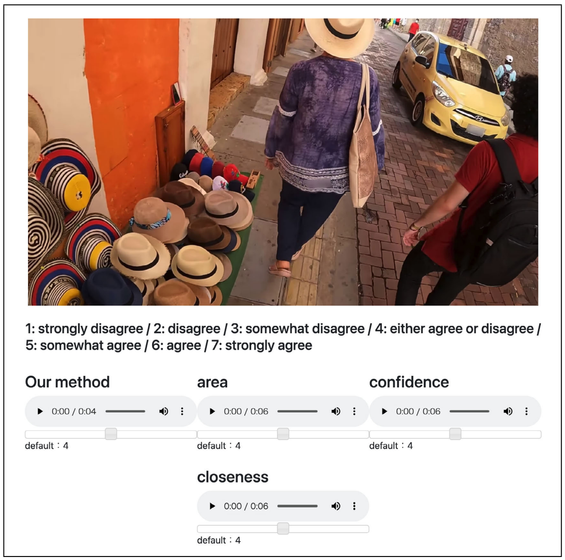
where $\mathbf{DCG@5_{pred}}$ and $\mathbf{DCG@5_{true}}$ are $\mathbf{DCG@5}$ of the estimated ranking and the ground-truth ranking, respectively. The output becomes a low value if the top rankings are incorrect since the high-ranked information is strongly weighted in the calculation of this metric.

C. RESULTS OF QUANTITATIVE EXPERIMENT

As shown in Table 2, our method reported higher estimation accuracy for all evaluation metrics than all baseline methods. For example, in NDCG@5, our method improved accuracy by 8.8% for area, 11.5% for confidence, and 20.7% for closeness. There was a significant improvement compared to the ranking estimation accuracy of the two methods (confidence and closeness). Although the ranking accuracy by area was higher than the other baselines, it was still lower than our method. The accuracy of NDCG@5 indicated that the estimated rankings from our method were more accurate overall than those from baselines. Furthermore, the results of Top-N Accuracy showed that our method was more likely to estimate a higher priority for the 1st-rank object than the baselines. Finally, from the results of Top-N Accuracy (Completely), our method estimated more accurately ranks the objects, not just the 1st-rank one.

D. QUALITATIVE EXPERIMENT

We conducted a qualitative evaluation by qualified caregivers who assist daily activities of VIB people. Four types of warning audio generated by the three baseline methods and our method were displayed and played for each image. The warning audio is augmented with information about the location of each object with the user. We asked four caregivers (three men and a woman) to give their subjective scores for 50 cases, in total we collected 200 responses. All the caregivers had a qualified status of a guide helper to support VIB people. The input image is divided horizontally into three sections, and the objects' position are defined as left, front, and right. For each object, the positional relationship was determined by which partition the center of the bounding box belonged. The participants were asked about the satisfaction of each audio using a slide bar. We instructed participants to rate the level of agreement with each audio based on whether the audio could prioritize warning of dangerous or critical objects during support for VIBs' walking. These were asked on a scale of 1 to 7, where 1 = strongly disagree and 7 = strongly agree. Each participant answered these Likert scale questions for 50 images. This experiment was conducted on a Web-based tool (Figure 4).



1: strongly disagree / 2: disagree / 3: somewhat disagree / 4: either agree or disagree / 5: somewhat agree / 6: agree / 7: strongly agree

Our method
0:00 / 0:04
default : 4

area
0:00 / 0:06
default : 4

confidence
0:00 / 0:06
default : 4

closeness
0:00 / 0:06
default : 4

FIGURE 4. Web-based tool for the qualitative evaluation.

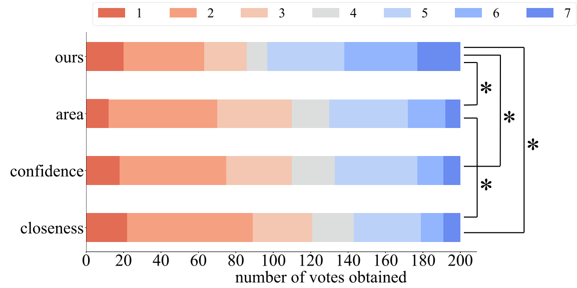


FIGURE 5. Result of the qualitative experiment *(*: significant effect)*.

E. RESULTS OF QUALITATIVE EXPERIMENT

Figure 5 illustrates the result of the qualitative experiment. A Friedman test indicated a significant difference in the level of agreement of these methods(χ^2=53.03, p<.05). Then we conducted post hoc analysis using Wilcoxon Signed Rank tests with Bonferroni correction (alpha=0.00167). There were significant differences(p<.0001) in all the tests related to our method. The results suggested that the participants highly preferred the warning audio with our method to all other baselines. It shows that the warning audio with our method could be agreeable for qualified caregivers of the VIBs.

VI. DISCUSSION

A. DISCUSSION OF EXPERIMENTAL RESULTS

In the quantitative experiment, especially in Top-N Accuracy, there were slight differences between area and our method.

	Ground-Truth	Ours	Area	Confidence	Closeness
1	Front bicycle	Front bicycle	Left pole	Left car	Right person
2	Left car	Left car	Left car	Front bicycle	Front person
3		Left pole	Front bicycle	Front person	Front bicycle
...			...	...	...
16			Right person	Left person	Left car
17			Left person	Left car	Left car

(a) Case example 1

	Ground-Truth	Ours	Area	Confidence	Closeness
1	Front pole	Front braille block	Front pole	Front person	Front braille block
2	Front braille block	Front pole	Front braille block	Left person	Front person
3	Left pole	Left pole	Left pole	Front person	Front person
4	Left tree		Left tree	Front braille block	Front pole
5	Front person		Front person	Front person	Front person
6			Left grass	Left tree	Front ballard
7			Front tree	Left pole	Front ballard
...			...	...	...
13			Front ballard	Front pole	Left tree
...			...	...	...
16			Front ballard	Front pole	Left pole

(b) Case example 2

FIGURE 6. Case examples of how warning audio navigates the category name of important objects. Ground-truth denotes the annotated objects' rankings. The objects ranked 1st, 2nd, and 3rd in ground-truth are shown in red, blue, and green, respectively. The warning audio from the three baselines has all the information on detected objects since they have not considered the warning threshold. The length of warning audio from baselines was 24 seconds and 23 seconds, respectively. In contrast, the audio from our method has only the information on the important objects. It took 6 seconds and 5 seconds, respectively, to warn with the audio from our method.

Therefore, we compared the ranking estimation results for those two methods. We found that area fails to estimate rankings of important objects with small areas (*e.g.*, bollards, safety_cones). In addition, area ranked cars highly even if they were parked and they did not disturb users to walk. In contrast, our method ranked the important small objects at the top rankings and ranked low for movable objects in stopping. However, neither area nor our method could estimate the rank accurately in scenes where all objects were far from the user.

In the qualitative experiment, the participants answered strongly agreeable in scenes where the object in front of the user was correctly estimated. In addition, even if rankings from multiple methods are similar, our method was evaluated well if our method was achieved to reduce warning objects significantly. On the other hand, the participants disagreed with our method in scenes where warning objects were excessively removed. In particular, the participants disagreed strongly in all scenes in which no objects were warned.

B. CASE EXAMPLES

Figure 6 shows two case examples of how warning audio navigates the category names of important objects. The objects ranked 1st, 2nd, and 3rd in ground-truth are shown in red, blue, and green, respectively. The method presented only the necessary information in the proper order, similar to ground-truth. Our method made much reduction of warning time by extracting important objects. The reduction in warning time helps to alleviate the effects of changes in the surrounding environment as the user walks.

In case example 1, our method estimated three of the 17 objects in the scene to be warning targets. These three objects included all objects that the annotator had identified as important. As a result, our method presented the presence of the highly important objects in six seconds, while it took 24 seconds to warn all detected objects. Since cars are large and have a high degree of danger when approaching, their warning orders generally tend to be high. However, in this scene, it is located far from the car user and is moving in the direction away from the user. Our method recognized the situation and ranked a bicycle higher than a car, unlike rankings by area or confidence. In area, a method that arranges the objects in order of area, the top ranking was a pole that was not ranked in the ground-truth. Therefore, the warnings of two important objects were delayed. In addition, unlike in our method, the ranking of cars is higher than that of bicycles. In confidence, which orders YOLOv5's detection confidence scores in descending order, objects to be warned were appropriately predicted to be higher. On the other hand, the rankings of person objects were higher than other objects. Especially in this scene, it was unclear which person was especially important because of the number of person objects. In closeness, which warns the users from nearby objects, the objects to be warned were estimated to be in the 3rd and 16th, respectively. As with confidence, it could not clarify which person was being warned.

In case example 2, Our method warned three of 16 objects in the scene. The warning time of our method was five seconds, whereas it needed 23 seconds to warn all detected objects. Although the order of the 1st and 2nd places was reversed, top-rank objects were well ranked. However in this scene, our method failed to warn 4th-rank and 5th-rank objects. Area was similar to our method regarding the warning order. In particular, the top-ranking objects were perfectly consistent with ground-truth. However, since the three baselines are supposed to warn all detected objects, the warning time was much longer than our method. In confidence, "person" objects are preferentially alerted as in case example 1. As we mentioned in Sec. V-B1, it was easier for YOLOv5 to detect person objects with higher confidence scores than other objects in urban scenes. In contrast, the top three objects in ground-truth were ranked low by this method (ranked 4th, 7th, and 13th, respectively). In closeness, the pole was not preferentially ranked even though it was just in front of the users. This pole indeed existed in front of the user. However, since the height of this pole is big, the distance between the center of the bounding box and the user was bigger than those of other objects. The overall ranking also significantly deviated from the ground-truth. Especially the Left pole, which was annotated in 3rd place in ground-truth, was ranked 16th in closeness.

There was a gap between the ground-truth and our method regarding the warning threshold for warning. In case example 1, the left pole, predicted as the 3rd rank, was not considered important in the ground-truth. In case example 2, our method should have warned two more objects. The results showed that the accuracy of the warning threshold needed to be higher. In particular, we need to improve the accuracy of the warning threshold in scenes as in case example 2, since an excessive reduction of warning objects may lead to the user overlooking the impending danger. However, our method outperforms the three baseline methods in significantly reducing the warning time with the high accuracy of ranking estimation. Our method gives us another chance to warn about impending dangerous objects.

VII. LIMITATIONS

We proposed a importance ranking estimation framework to provide walking assistance to the VIBs. The experiments showed that our ranking estimation method outperforms other baselines. However, several issues need to be resolved regarding the practical use of this method in the real world.

A. RANKING ESTIMATION IN VIDEOS

To use our method in the real world, we need to consider the transition of users' walking state. To accomplish this, it is necessary to detect and rank objects with videos. We need to map objects in each frame to the frames before and after it. In other words, we must keep track of the detected objects. We can achieve this requirement by tracking each detected object using models for multi-object tracking [41], [42], [43].

B. COLLECTING DATA FROM VIB INDIVIDUALS

There is a difficulty of collecting data from VIB individuals. Since VIB individuals cannot obtain visual information about their surroundings, it is difficult to ask them how dangerous each object is. Furthermore, asking them how often to be warned is equally difficult. We also need to consider continuous changes around the user. As mentioned above, the dataset used in this paper was annotated by a qualified caregiver for the visually impaired. We believe annotators who can use visual information in this way can generate alternative datasets. For example, we can collect a dataset that satisfies our purpose by synchronously capturing the caregiver's voice and the VIBs' first-person video.

C. FURTHER DETECTION OF IMPORTANT ELEMENTS OF THE ROAD

When assisting VIB individuals, detecting additional types of objects on the road might be helpful for effective navigation. For example, detecting objects under the VIB users' feet (*E.g.*, steps, puddles, white lines) might be useful for assessing the safety of walking. In fact, several participants suggested this point during the quantitative experiment. In particular, some participants stated that the information about the white line should be included in the object warning system like this research. Since no step is on the white line, it is very difficult for VIBs to know themselves on the road or the sidewalk.

D. OPTIMIZING WARNING AUDIO

Another participant stated, "I could not understand what was most dangerous just by reading it out loud with flat intonation." Stated differently, a strong tone of machine voice may be necessary when danger is imminent. We also need to consider a new type of warning audio. For example, it is important to inform the user of not only the warning of white lines but also the information on whether the user is on the road or the sidewalk.

From a different perspective, our study adopted a certain reading speed and did not optimize the reading speed of machine voices. People using machine speech since childhood can cope with faster speech [44]. Therefore, optimization of playback speed should be done for each user.

VIII. CONCLUSION

We proposed an object importance estimation method for walking support for VIB people. Within this method, we proposed a new frame for importance ranking estimation. We used optical flow, including time series information, which enabled us to estimate rankings based on the state of the movable objects. Using the relaxed permutation matrices with NeuralSort, the network was trained directly with the order. The proposed method accurately estimated the rankings compared to the three baseline methods. Additionally, the people assisting the VIBs agreed on the warning audio with our method. Our method is expected to be applied in many applications related to walking assistance. In the future, we will improve the detection object and warning method based on the opinions given by caregivers in the quantitative experiments. From a different point of view, we will develop a warning system for sequences, considering the transition of users' walking state.

APPENDIX A
NeuralSort

This appendix provides an overview of NeuralSort. NeuralSort is a differentiable sorting method that can train networks with the original ranking form. As a premise, rankings are expressed with permutation matrices in differentiable sorting methods. Specifically, the permutation matrix $P_{\mathbf{z}}[i, j]$ is given by:

$$P_{\mathbf{z}}[i,j] = \begin{cases} 1 & \text{if } j = z_i \\ 0 & \text{otherwise,} \end{cases} \qquad (10)$$

where $\mathbf{z} = [z_1, z_2, \ldots, z_n]^T$ is an n-dimensional permutation of a list of unique indices $\{1, 2, \ldots, n\}$. These matrices allowed us to directly train models with the ranking and not consider the ranking scores' distribution.

However, the permutation matrices like Eq. 10 are non-differentiable since the argmax operator generates those matrices. To train networks regarding rankings, relaxed permutation matrices are used in NeuralSort. The i-th row of the relaxed permutation matrices ($\widehat{P}_{sort(\mathbf{s})}[i, :](\tau)$) is given by:

$$\widehat{P}_{sort(\mathbf{s})}[i, :](\tau) = \text{softmax}\left[\frac{(n+1-2i)\mathbf{s} - A_{\mathbf{s}}\mathbf{1}}{\tau}\right], \qquad (11)$$

where $\mathbf{s} = [s_1, s_2, \ldots, s_n]^T$ is the matrix of ranking scores. In other words, s_i denotes the importance score of i-th information or object. $sort(\mathbf{s})$ is the sorting operation of $\mathbf{s}$ whose indices are assigned in order. For example, if the ranking score vector $\mathbf{s} = [3, 6, 2]^T$, $sort(\mathbf{s}) = [2, 1, 3]$, since the top-ranking score is at the 2nd index, the 2nd-biggest ranking score is at the 1st index, and so on. $A_{\mathbf{s}}[i, j] = |s_i - s_j|$ denotes the matrix of absolute pairwise differences between s_i and s_j. $\mathbf{1}$ denotes the column vector of all ones. τ is a temperature parameter that controls the degree of smoothness of the relaxed permutation matrix $\widehat{P}_{sort(\mathbf{s})}$, and $P_{sort(\mathbf{s})}$ is consistent with $\widehat{P}_{sort(\mathbf{s})}$ at $\tau \rightarrow 0+$. We calculated this equation with $\tau = 1$. The proof of Eq. 11 is given in COROLLARY 3 of NeuralSort. The relaxed permutation matrices are continuous and differentiable since these matrices are calculated by the softmax operator.

APPENDIX B
URLs OF YouTube VIDEOS

Here are the URLs of YouTube videos that we used to construct the importance ranking dataset.

- https://www.youtube.com/watch?v=gH9Zf-AbykA. 20 Nov 2022
- https://www.youtube.com/watch?v=E31MMNBpw_g. 20 Nov 2022
- https://www.youtube.com/watch?v=jxrCF8uN79M. 20 Nov 2022
- https://www.youtube.com/watch?v=w-3p-OOhGx4. 20 Nov 2022
- https://www.youtube.com/watch?v=kDZpAWmn2J8. 20 Nov 2022

APPENDIX C
LIST OF ACRONYMS

Following acronyms were used in this paper.

TABLE 3. **List of acronyms.**

Full name	Acronym
Visually Impaired/Blind	VIB
Learning-to-Rank	LTR
Support Vector Machine	SVM
Discounted Cumulative Gain	DCG
Normalized Discounted Cumulative Gain	NDCG

REFERENCES

[1] R. Vaz, D. Freitas, and A. Coelho, "Blind and visually impaired visitors' experiences in museums: Increasing accessibility through assistive technologies," *Int. J. Inclusive Museum*, vol. 13, no. 2, pp. 57–80, 2020.

[2] M. R. Ahmed and M. A. Naveed, "Information accessibility for visually impaired students," *Pakistan J. Inf. Manage. Libraries*, vol. 22, pp. 16–36, Dec. 2020.

[3] 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, Apr. 2019, Art. no. 044102.

[4] L. Zeng, D. Prescher, and G. Weber, "Exploration and avoidance of surrounding obstacles for the visually impaired," in *Proc. 14th Int. ACM SIGACCESS Conf. Comput. Accessibility*, Oct. 2012, pp. 111–118.

[5] M. A. Williams, C. Galbraith, S. K. Kane, and A. Hurst, "Just let the cane hit it': How the blind and sighted see navigation differently," in *Proc. 16th Int. ACM SIGACCESS Conf. Comput. Accessibility (ASSETS)*, 2014, pp. 217–224.

[6] Md. M. Islam, M. Sheikh Sadi, K. Z. Zamli, and Md. M. Ahmed, "Developing walking assistants for visually impaired people: A review," *IEEE Sensors J.*, vol. 19, no. 8, pp. 2814–2828, Apr. 2019.

[7] M. Shimakawa, K. Matsushita, I. Taguchi, C. Okuma, and K. Kiyota, "Smartphone apps of obstacle detection for visually impaired and its evaluation," in *Proc. 7th ACIS Int. Conf. Appl. Comput. Inf. Technol.* New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–6.

[8] N. Rachburee and W. Punlumjeak, "An assistive model of obstacle detection based on deep learning: Yolov3 for visually impaired people," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, pp. 3434–3442, 2021.

[9] K. C. Shahira, S. Tripathy, and A. Lijiya, "Obstacle detection, depth estimation and warning system for visually impaired people," in *Proc. IEEE Region 10 Conf. (TENCON)*, Oct. 2019, pp. 863–868.

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

[11] A. Grover, E. Wang, A. Zweig, and S. Ermon, "Stochastic optimization of sorting networks via continuous relaxations," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–23.

[12] P. Strumillo, "Electronic interfaces aiding the visually impaired in environmental access, mobility and navigation," in *Proc. 3rd Int. Conf. Human Syst. Interact.*, May 2010, pp. 17–24.

[13] B. B. Blasch, W. R. Wiener, and R. L. Welsh, *Foundations of Orientation and Mobility*, 2nd ed. New York, NY, USA: AFB Press, 1997.

[14] S. Tachi, "Guide dog robot," in *Proc. 2nd Int. Symp. Robot. Res.*, 1984, pp. 333–340.

[15] J. K. Mahendran, D. T. Barry, A. K. Nivedha, and S. M. Bhandarkar, "Computer vision-based assistance system for the visually impaired using mobile edge artificial intelligence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2418–2427.

[16] M. A. Khan, P. Paul, M. Rashid, M. Hossain, and M. A. R. Ahad, "An AI-based visual aid with integrated reading assistant for the completely blind," *IEEE Trans. Hum.-Mach. Syst.*, vol. 50, no. 6, pp. 507–517, Dec. 2020.

[17] S. Caraiman, A. Morar, M. Owczarek, A. Burlacu, D. Rzeszotarski, N. Botezatu, P. Herghelegiu, F. Moldoveanu, P. Strumillo, and A. Moldoveanu, "Computer vision for the visually impaired: The sound of vision system," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1–9.

[18] Y. Chang, M. Kutbi, N. Agadakos, B. Sun, and P. Mordohai, "A shared autonomy approach for wheelchair navigation based on learned user preferences," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1490–1499.

[19] M. Inaba and K. Takahashi, "Neural utterance ranking model for conversational dialogue systems," in *Proc. 17th Annu. Meeting Special Interest Group Discourse Dialogue*, 2016, pp. 393–403.

[20] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retr.*, vol. 3, no. 3, pp. 225–331, 2009.

[21] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 129–136.

[22] M. Taylor, J. Guiver, S. Robertson, and T. Minka, "SoftRank: Optimizing non-smooth rank metrics," in *Proc. Int. Conf. Web search Web Data Mining (WSDM)*, 2008, pp. 77–86.

[23] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 271–278.

[24] J. Xu and H. Li, "AdaRank: A boosting algorithm for information retrieval," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2007, pp. 391–398.

[25] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 89–96.

[26] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.

[27] C. Burges, R. Ragno, and Q. Le, "Learning to rank with nonsmooth cost functions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2006, pp. 1–8.

[28] M. Cuturi, O. Teboul, and J.-P. Vert, "Differentiable ranking and sorting using optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[29] R. Swezey, A. Grover, B. Charron, and S. Ermon, "PiRank: Scalable learning to rank via differentiable sorting," in *Advances in Neural Information Processing Systems*, vol. 34. Red Hook, NY, USA: Curran Associates, 2021, pp. 21644–21654.

[30] M. Blondel, O. Teboul, Q. Berthet, and J. Djolonga, "Fast differentiable sorting and ranking," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 950–959.

[31] S. Prillo and J. Eisenschlos, "SoftSort: A continuous relaxation for the argsort operator," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 7793–7802.

[32] F. Petersen, H. Kuehne, C. Borgelt, and O. Deussen, "Differentiable top-K classification learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 17656–17668.

[33] X. Liu, C. Yu, Z. Zhang, Z. Zheng, Y. Rong, H. Lv, D. Huo, Y. Wang, D. Chen, J. Xu, F. Wu, G. Chen, and X. Zhu, "Neural auction: End-to-end learning of auction mechanisms for e-commerce advertising," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 3354–3364.

[34] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scand. Conf. Image Anal.* Cham, Switzerland: Springer, 2003, pp. 363–370.

[35] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. J. Conf. Artif. Intell.*, vol. 2, Aug. 1981, pp. 674–679.

[36] G. Jocher, *YOLOV5 by Ultralytics*. Accessed: Nov. 20, 2022. [Online]. Available: https://github.com/ultralytics/yolov5

[37] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[38] W. Tang, D.-E. Liu, X. Zhao, Z. Chen, and C. Zhao, "A dataset for the recognition of obstacles on blind sidewalk," *Universal Access Inf. Soc.*, vol. 22, pp. 69–82, Aug. 2021.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[40] K. Grauman, "Ego4D: Around the world in 3,000 hours of egocentric video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18973–18990.

[41] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*.

[42] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3069–3087, Nov. 2021.

[43] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "TransMOT: Spatial-temporal graph transformer for multiple object tracking," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4859–4869.

[44] D. Bragg, C. Bennett, K. Reinecke, and R. Ladner, "A large inclusive study of human listening rates," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2018, pp. 1–12.

YASUHIRO NITTA (Graduate Student Member, IEEE) received the B.E. degree in information and computer science from Keio University, Japan, in 2022, where he is currently pursuing the M.S. degree in science and technology. His research interests include assistance for visually impaired and importance ranking estimation based on people's subjectivity.

MARIKO ISOGAWA (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Osaka University, Japan, in 2011, 2013, and 2019, respectively. From 2019 to 2020, she was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA. She is currently an Associate Professor with the Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Japan. Her research interests include computer vision and pattern recognition.

RYO YONETANI (Member, IEEE) received the B.E., M.S., and Ph.D. degrees from Kyoto University, Japan, in 2009, 2011, and 2013, respectively. He is currently a Project Senior Assistant Professor with Keio University, Japan. His research interests include machine learning and computer vision.

MAKI SUGIMOTO (Member, IEEE) received the Ph.D. degree from The University of Electro-Communications, Tokyo, in 2006. He is currently a Professor with the Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Japan. His research interests include embedded optical sensing systems for AR/VR and wearable sensing systems with machine intelligence.