Human-like Guidance with Gaze Estimation and Classification-based Text Generation

Masaki Nambata[1] Kota Shimomura[1] Tsubasa Hirakawa[1] Takayoshi Yamashita[1] Hironobu Fujiyoshi[1]

Abstract— Car navigation systems are widely used and essential for driving assistance. However, drivers often struggle to understand voice guidance from these systems, leading to the need for constant map-checking on a monitor, which can be dangerous. In contrast, human guidance utilizing visible objects is clearer to drivers. Human-like Guidance (HℓG) is a task that realizes such human-like navigation on a system. In this paper, we propose a novel method for HℓG. Our approach involves defining human-like navigation templates and selecting appropriate sentences for each object in an intersection scene. We also construct a model to estimate the driver's gaze and use this information to choose a reference object for navigation, resulting in a system that provides clear guidance to the driver. Furthermore, we provide a gaze information dataset called the Driving Gaze Dataset to build a driver gaze estimation model. Through experiments using the CARLA automated driving simulator, we demonstrate the feasibility of generating navigation instructions that drivers can intuitively understand. In addition, we confirmed that our method is able to generate navigation quickly. This research is expected to mitigate risky driving caused by navigation systems while driving.

I. INTRODUCTION

Car navigation systems can be categorized into two types: automated car navigation and human-guided navigation. Automated car navigation systems, which are commonly used today, utilize GPS and digital maps to provide guidance through pre-formatted sentences displayed on a monitor.However, these instructions can be difficult to understand intuitively and may lead to misinterpretations and distractions for drivers. Human-guided navigation, on the other hand, relies on information from the surrounding environment, providing situation-based and easily understandable guidance. This approach takes into account landmarks and nearby objects, reducing the cognitive load on drivers and minimizing the risk of navigation errors. It is worth noting that human-guided navigation requires the presence of passengers to provide guidance.

In this paper, our objective is to realize Human-like Guidance (HℓG), which aims to replicate human-like navigation in a system. HℓG enables the system to independently generate guidance text on the basis of the situation, resembling human guidance. For example, the system might generate instructions such as "Follow the red car ahead and make a right turn at the next intersection."

[1]Authors are with Chubu University, 1200 Matsumotocho, Kasugai, Aichi, Japan, 487-8501 masaknanbt@mprg.cs.chubu.ac.jp

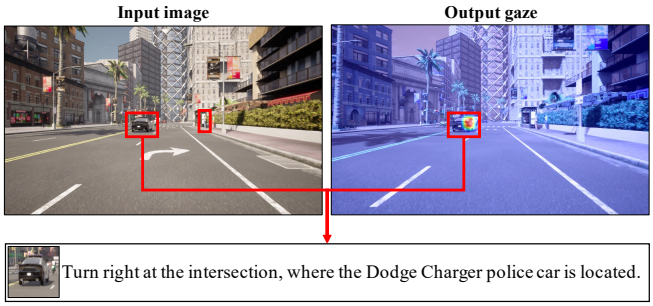


Fig. 1. Overview of our proposed Human-like Guidance (HℓG) method.

Previous research has explored HℓG as a means of augmenting existing car navigation systems, as demonstrated in Apple's patent [1], or as a method utilizing computer vision, as demonstrated by Bihao et al. [2]. Apple's patented approach involves incorporating information about buildings, such as restaurants, onto a map, thus generating instructions like "Turn left in 100 meters The restaurant is on the left." However, relying solely on map-based information poses the risk of potential inaccuracies and outdated information, which could mislead drivers. Bihao et al.'s proposed method addresses this issue by accurately placing buildings and other objects detected from in-vehicle camera images onto an aerial map.

However, if the object selected as a navigation criterion differs from the object perceived by the driver, the driver will need to search for the specified object, potentially increasing their risk. To address this issue, this study proposes a method for achieving HℓG by incorporating information on the driver's gaze during driving. We show an overview of our proposed method in Figure 1. We pre-define human-like navigation sentences based on objects and use a classification model to determine the appropriate sentences corresponding to the objects observed by the in-vehicle camera. Subsequently, a gaze estimation network is utilized to gather information about the driver's gaze while driving, enabling the system to select a navigation reference that is easily understandable for the driver on the basis of the objects estimated to be the focus of the driver's attention.

However, determining the precise object to be used as a navigation reference presents a challenge due to the rapid changes in the human driver's line of sight during driving. Therefore, this study proposes the development

of a gaze dataset during driving, called the Driving Gaze Dataset (DGD), specifically designed for the HℓG task. This dataset facilitates more accurate and optimal object selection within a given scene.

The contributions of this research are as follows:

- We propose a novel baseline method for HℓG, addressing the real-time and safety concerns associated with current car navigation systems.
- We introduce a gaze information dataset specifically designed for HℓG, enabling the utilization of gaze information in the navigation process.

II. RELATED WORKS

The HℓG task we are working on has, as far as we can tell, very little prior research, and we are the first to generate human-like car navigation. In this section, we describe image captioning and gaze estimation methods. Therefore, in this section, we describe image captioning and gaze estimation methods related to our method.

A. Image captioning

One of the research topic which is related with our study is image captioning in the field of vision and language. Image captioning generates a descriptive text for a given image. Image captioning mainly uses a model that consists of two steps: i) a feature extraction step from an input image and ii) generating a description from the extracted features.

Image captioning have been studied before the advent of deep learning. A classical method typically estimates word labels for the whole image or each object region through image identification as the first step. Then, we can generate a caption by fitting the words obtained from the estimated labels to a prepared perforated template in it [3].

Since the deep learning era, the most of methods tend to use CNNs to extract image features and RNNs to generate an explanatory sentence [4], [5] After the probability of a highly accurate object detection method, a method combining feature extraction by CNN and object detection has been proposed Neural Baby Talk (NBT) [6] introduced a CNN-based object detection method. NBT has developed the method [3] using template captions with holes in the past, and has also proposed a method to automatically generate template captions with holes in the image corresponding to the image.

After the appearance of Transformer [7], various methods of using Transformer for caption generation have been proposed. Simao et al. achieved highly accurate caption generation using Transformer, which considers the positional relationship for each object detected by object detection [8]. GRIT [9] achieves high accuracy and high speed at the same time by building a network that integrates Grid features and Region features using only Transformer.

In our study, it is important to generate human-like navigation with reference to objects on the input scene. However, learning specialized sentences such as H*ell*G using natural language generative models such as RNNs and Transformer is difficult due to the large datasets required and workload. Furthermore, since the instructions for car navigation are only needed in limited situations, such as at intersections, it is possible to limit the number of instruction sentences to a few types, eliminating the need to use a natural language generation model. Therefore, we define a human-like car navigation template and perform sentence generation by solving the decision of the template as a classification task. In our method, appropriate navigation templates for objects in the input image are used as ground truth data and trained. The appropriate templates change depending on the type and location of the object. Apart from the aforementioned image captioning methods, our approach classifies templates for all objects at the intersection.

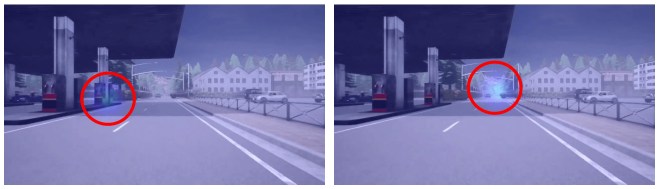
B. Gaze prediction

Gaze estimation is the task of predicting where a human will look given an image or video. The task of algorithmically reproducing gaze estimation is called visual saliency prediction. Many visual saliency prediction methods have been proposed and developed using CNN and deep learning. Ensemble of Deep Networks (eDN) proposed by Vig et al. is the first model that uses CNN [10]. After that, models such as AlexNet [11] and VGGNet [12] were successfully used as feature extractors to perform saliency prediction, and this became a common form of saliency prediction [13], [14]. One of the representative methods for visual saliency prediction is the Dilated Inception Network (DINet) [15]. In DINet, the Dilated Residual Network (DRN) [16] is used as the encoder, and after capturing multi-scale features with Dilated Inception Modules (DIM), the decoder generates a saliency map.

DRN is a network that introduces dilated convolutions to ResNet-50 [17]. DRN is a pre-trained network often used for visual saliency prediction.

DIM introduces a dilated convolution process to the Inception module proposed in GoogLeNet [18]. By replacing each convolutional process in the Inception module with a dilated convolution process with a different dilation rate, it is possible to achieve the same or better accuracy than existing methods with a small number of parameters and short training time.

In the realization of HℓG, it is necessary to select a navigation reference object that is easy for drivers to understand, but there is no clear definition of a navigation reference object. Therefore, this study uses a gaze estimation model to estimate the driver's gaze and to select objects as navigation references on the basis of the model's estimated gaze.



(a) With instructions (b) Without instructions

Fig. 2. Gaze data examples of scene before left turn. For the ease of clarity, we show the position of gaze data as a red circle. This example shows that a participants gazes at gas station after receiving instructions.

TABLE I

Scene annotation definitions

Scene	Label	Number of frames
Following traffic lane	0	Always
Straight	1	240 frames from stop line
Turn right	2	Until vehicle is in vertical position from stop line
Turn left	3	Until vehicle is in vertical position from stop line
Stop	4	From vehicle stop to start
Accidents	5	From lane entry to frame out
Before rigth turn	6	300 frames before stop line
Before left turn	7	300 frames before stop line

III. DRIVING GAZE DATASET

Several datasets have been developed to collect human gaze information during car driving [19], [20]. When driving, it is essential to visually attend to various aspects of the road, resulting in rapid changes in human gaze. However, for the purpose of selecting the most suitable object for navigation in the context of HℓG utilizing gaze information, rapidly changing gaze data is not be ideal. Hence, we developed an optimal dataset, called the Driving Gaze Dataset (DGD), specifically designed for HℓG using gaze information. The DGD consists of three types of data: driving video data, gaze information data, and scene annotation data. Hereafter, we provide a detailed descriptions on the DGD.

A. Driving video data

Obtaining driving video data containing a wide range of real-world scenes can be challenging due to the associated workload. To address this, we utilized CARLA, a development simulator for automated driving, to generate data from the driver's perspective [21].

We created a dataset consisting of 20 videos, each approximately 3 minutes long, captured at 60 frames per second (fps). Furthermore, we collected data in a manner that ensured every frame of the driving video data was associated with gaze information data and scene annotation data. The gaze information data was collected by recording the driver's gaze while viewing the driving video data. The scene annotation data was annotated with a class label assigned to each frame that represents the driving scene.

B. Gaze information data

Gaze data was collected using the Tobii-Pro X3-120 eye-tracking device, which recorded eye movements while participants viewed videos generated by the CARLA simulator. Within the scene leading up to the intersection, participants were instructed to direct their gaze towards objects that could serve as navigation references, including vehicles making turns ahead, pedestrians at the intersection, and prominent buildings. An illustrative example comparing data collected with and without navigation instructions is depicted in Figure 2. To account for individual variations in gaze patterns, data was collected from a total of eight subjects as part of this study.

C. Scene annotation data

We classified various potential driving scenarios, such as driving straight, making a right turn, making a left turn, etc., into eight distinct categories. Table I shows scene categories for the DGD we defined. This dataset provides the gaze estimation model with crucial information regarding the specific driving situation and the intended travel direction within the input scene.

IV. PROPOSED METHOD

As mentioned, our objective is to achieve HℓG, that is, human-like navigation that relies on scene objects as references. However, existing natural language processing algorithms and datasets do not provide human-like navigation specifically tailored to complex scenarios like intersections. Furthermore, there is no definitive definition of what constitutes an ideal reference for human-like navigation. To address these challenges, we propose a novel approach for HℓG that uses a method that classifies navigation templates for each object using image classification techniques and selects the most suitable object through gaze estimation.

An overview of the proposed method is depicted in Figure 3. This approach involves multi-class classification, utilizing images cropped by a bounding box (BBox) of all objects within the input scene, along with the distance between the object and the intersection. Additionally, the same scene is inputted into the gaze estimation model trained with the DGD to estimate the driver's gaze. The model then selects an object as the reference for navigation on the basis of the gaze estimation and generates navigation sentences corresponding to the class to which the object belongs.

A. Navigation template selection as classification

In this study, we generate navigation sentences that incorporate specific target objects, such as "Turn left at the intersection, where the Dodge Charger police car is located." This navigation sentence consists of two components: i) a navigation part indicating the driving direction of the vehicle in the first half and ii)

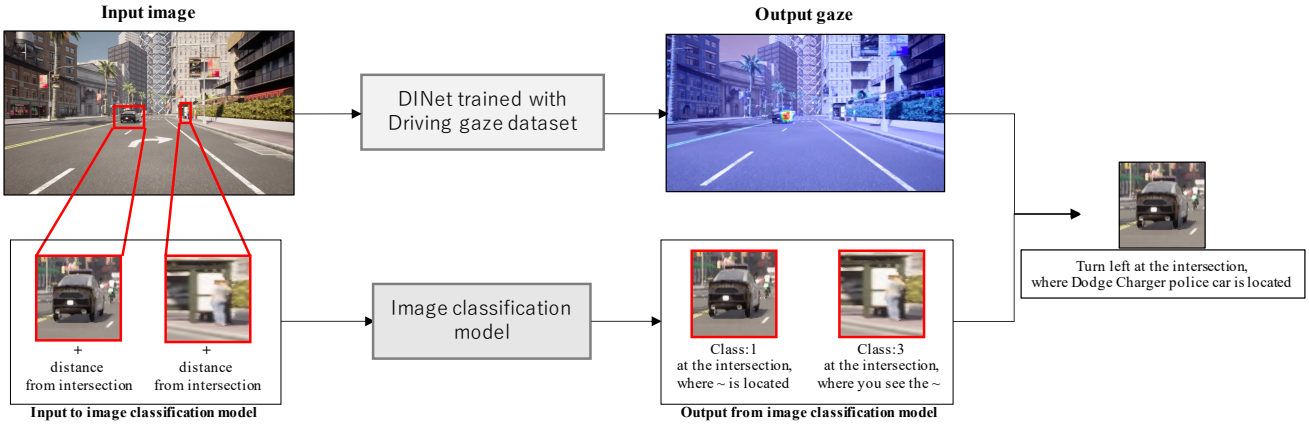


Fig. 3. Overview our proposed method

TABLE II
Navigation class definitions

Object	Distance from intersection (d)	Navigation template	Class
Car	20 >= d	at the intersection, following the	0
	20 <d	at the intersection, where the ~is located	1
Human	20 >= d	at the intersection, where you see the ~is nearby	2
	20 <d	at the intersection, where you see the ~is located	3
Others	None	at the intersection, where the ~is located	4

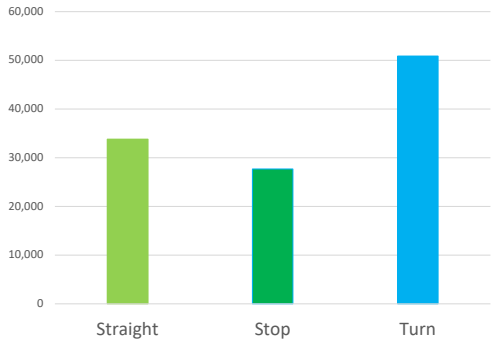


Fig. 4. Class ratio of input data

a reference part that identifies a target object in the latter half of the sentence. Moreover, the structure of the navigation sentence should be varied depending on the location of the target object because human changes the instruction depending on the object location adaptively. For example, if the target car is near an intersection, the sentence should be "Turn left at the intersection, where the car is located." If the target car is a bit farther from the intersection, the sentence should be "Turn left at the intersection, following the car." Therefore, we generate navigation sentences on the basis of a target object and the distance between the object and the intersection.

To accomplish this, we solve a classification task, as illustrated in the lower section of Figure 3. We use several pre-defined navigation templates, as shown in Table II. By classifying an object image with accompanying distance information, we determine the most appropriate navigation template.

Specifically, given BBox information, we extract object images from in-vehicle camera images. Subsequently, we feed these cropped object images, along with the distance information between the target object and the intersection, into a classification model. The model outputs classification probabilities for various navigation templates, and we select the template with the highest score. By adding the driving direction of the vehicle and the object information provided by CARLA to the selected templates, navigation is completed. This classification process is repeated for each candidate object in the scene, resulting in navigation assigned to individual objects.

B. Selecting optimal target object by gaze estimation

As mentioned previously, we generate navigation templates for various target objects, and we select the most suitable one to generate a human-like navigation sentence. To aid in this selection process, we leverage human gaze information. To replicate human gaze behavior during driving, we use DINet, a lightweight and computationally efficient gaze estimation model introduced in [15]. DINet takes an RGB image as input and produces a heatmap representing the estimated gaze location. During training, we calculate the loss by comparing the output heatmap with the ground truth gaze data. For the loss function, we adopt the linear normalization function proposed by Sheng et al. [15]. This function normalizes both the estimated gaze and the ground truth gaze, treating them as probability distributions. This normalization allows for the consideration of pixel-level relationships during the calculation.

During the reference object selection process, we utilize the DINet model trained with the DGD. However, since the DGD collects data randomly by autonomous driving,

(a) Right turn scene

(b) Straight scene

Fig. 5. Example scene at same intersection

there is a problem in the distribution of low-importance scenes (e.g., stops and straight driving) versus high-importance scenes (e.g., turns and scenes preceding turns) is almost equal (see Figure 4). To address this, we categorize the input data into three categories based on annotation labels: Straight (label: 0/1), Stop (label: 4/5), and Turn (label: 2/3/6/7). We adjust the data ratio to achieve a balanced distribution of 1:1:8, ensuring that the training process captures gaze behavior suitable for HℓG in crucial navigation scenes, such as intersections.

The object with the highest heatmap value within its bounding box (BBox) is selected as the optimal reference object on the basis of the heatmap output from DINet. This enables the system to choose the most appropriate object as a navigation reference in intersection scenes where multiple objects are present.

V. EXPERIMENTS

In this section, we conduct experiments in terms of the i) accuracy in navigation template classification for each target object and ii) the efficacy of selecting the best object by gaze estimation. Again, the purpose of this paper is to validate the efficacy of the baseline method, which is a new approach to HℓG. Therefore, we deal with a less difficult task. There are two verification methods as follows:

1) We conduct experiments on classifying objects using several image classification models, using as input an image of an object cropped to the size of the BBox and the distance of the object from the intersection.
2) We qualitatively compare the navigation with the object selected using the results of gaze estimation and the navigation with the object selected with the highest class probability in the scene.

A. Datasets

In our experiments, we utilized the proposed DGD introduced in SectionIII, along with another dataset collected using CARLA (referred to as the CARLA dataset). The DGD is used for training the gaze estimation model, while the CARLA dataset is utilized as the training data for the image classification model and as the evaluation data for the overall method in our experiments. Herein, we provide a brief overview of the CARLA dataset.

We generate intersection scenes using the CARLA simulator, creating four distinct scenes, each featuring a single intersection. (1. large intersection with two lanes in each direction, 2. T-intersection, 3. narrow intersection with no signal, 4. intersection with few buildings around) Within each intersection scene, we vary the vehicle's direction and the surrounding environment to generate diverse data samples. An example of the dataset is illustrated in Figure 5. Furthermore, since obtaining the coordinates of intersections and objects in 3D space directly from in-vehicle camera images is challenging, we leverage the functionality of CARLA to retrieve the coordinates of objects and intersections, enabling us to calculate the distances accurately.

TABLE III
Classification result

Model	Parameters	Inference time (ms)	Accuracy
ResNet-18	12,250,469	67.2	0.9817
VGG-16	138,985,285	71.0	0.9862
DenseNet	8,532,869	69.5	0.9863
MobileNet-v3	6,040,075	68.5	0.9902
DeiT-small-16	22,589,317	72.0	0.9928
MobileViT-v2	8,022,886	71.7	0.9860

B. Navigation template classification per object

First, we evaluated the accuracy of the navigation template classification task. As a classification model, we verify both CNN and Vision Transformer (ViT) [22] that have been pre-trained on ImageNet [23]. In our experiments, we performed a 5-class classification task. All cropped images, obtained using BBox, were resized to 224×224 pixels and utilized as input images for the classification models. For the CNN-based models, we used well-known architectures such as ResNet-18 [17], VGG-16 [12], DenseNet [24], and a more lightweight model, MobileNet-v3 [25], suitable for real-world implementation. Similarly, for the ViT-based models, we utilized representative architectures such as DeiT [26] and a smaller model, MobileViT-v2 [27], which is optimized for real-world implementation, akin to the CNN-based models.

Table III shows the results of the classifications for each model. Inference time in Table III indicates the inference speed per a single input. We measure the inference speed 5 times and take the average for each model. The results show that each models achieved a higher classification accuracy. It was confirmed that class classification was possible on the basis of only the input object images and features consisting of the absolute distance from the intersection.

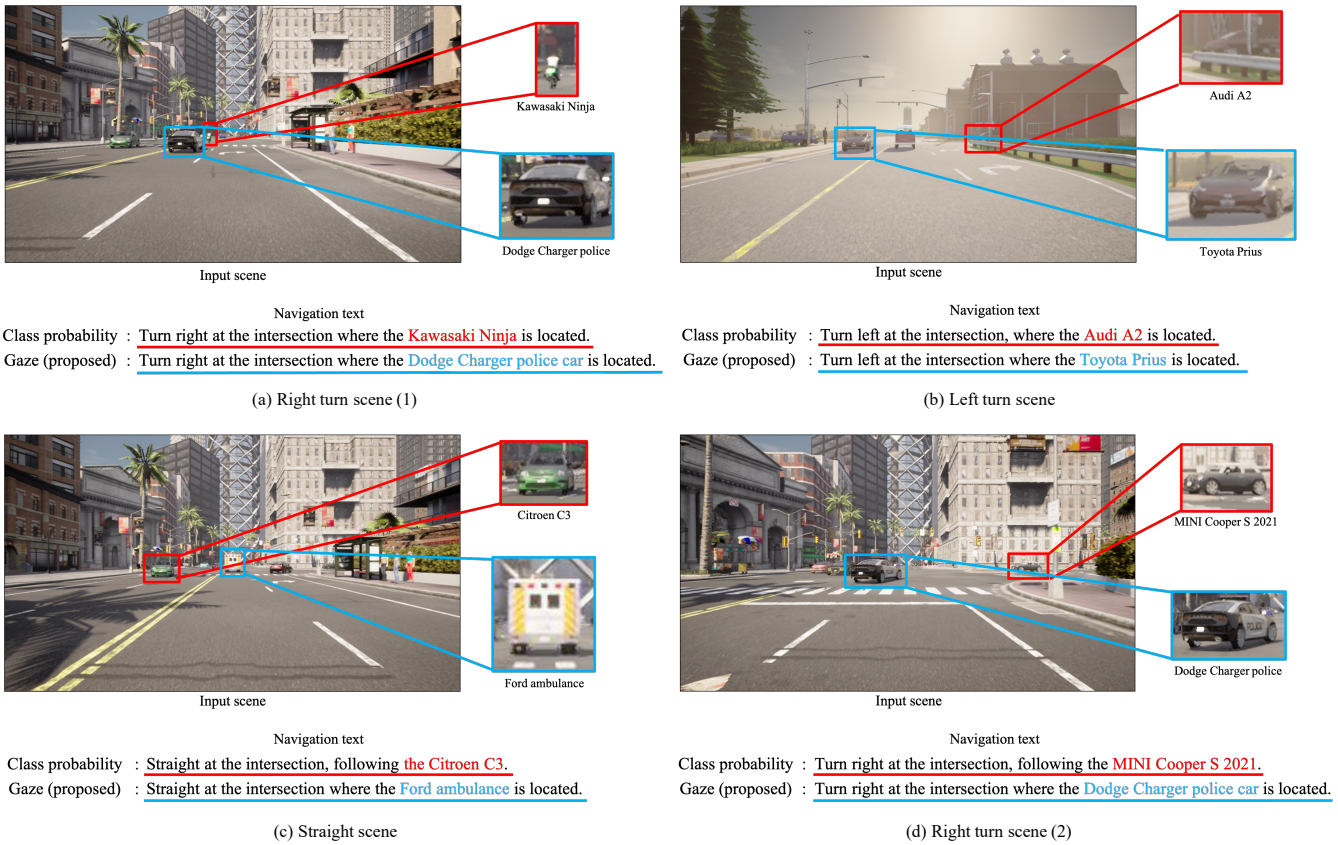


Fig. 6. Examples of output results.Objects bounded in red are selected with class probability. Objects bounded in blue are selected with predicted gaze.

C. Selecting reference object by gaze estimation

Next, we evaluated the accuracy of gaze estimation. To verify the effectiveness of selecting objects using gaze estimation, we qualitatively compared the navigation sentences generated by the object with the highest class probability in class classification with those generated by the object selected using gaze estimation. For gaze estimation, we utilized DINet, which was pre-trained with DGD, by inputting the same image data used for class classification. Regarding the classification results, we considered the outputs of ResNet-18, which offers the fastest computation speed and is suitable for real-world implementation. Experimental results confirm that our method can generate navigation in an average of 0.077 seconds per frame.

Figure 6 illustrates input intersection scene images and the resulting navigation statements. In the example shown in Figure 6(a), the object selected on the basis of class probabilities was a motorcycle located far from the driver. Conversely, the object selected using gaze estimation was a vehicle positioned in front of the driver. Similarly, in the example depicted in Figure 6(b), the object chosen from class probabilities was a vehicle behind a guardrail, whereas the object identified through gaze estimation was a vehicle in the oncoming lane at the intersection. These results demonstrate that gaze estimation enables the selection of objects that are easily understandable to drivers.

In the example presented in Figure 6(c), the object selected on the basis of class probabilities was a vehicle in the oncoming lane. However, as the classification is limited to five classes, it can be observed that this instruction was incorrect for following oncoming vehicles. In contrast, the object chosen via gaze estimation was a vehicle in the traveling direction. The output provided the correct instruction to proceed straight at the intersection where there was a vehicle in front. Similarly, in the example showcased in Figure 6(d), the object selected on the basis of class probabilities was a vehicle in the traveling direction. As the system generated navigation sentences for all objects in the input scene, it becomes apparent that the instructions were incorrect for following a car in the orthogonal lane. Conversely, the object selected using gaze estimation was the vehicle in front of the driver's own vehicle, resulting in a navigation instruction based on an object within the intersection, which confirms its correctness. These outcomes demonstrate that gaze estimation helps avoid the selection of objects that produce inaccurate instructions.

VI. CONCLUSION and DISCUSSION

In this paper, we proposed a novel approach to achieving Human-like Guidance (HℓG) in car navigation systems. Our approach combines class classification and gaze estimation to select objects as references for navigation. We introduced a dataset of gaze information (DGD) specifically designed for HℓG, which was proven to be optimal for our method. The experimental results show that our approach achieved high accuracy in terms of a navigation template classification task by using object image patch and distance information. In a gaze estimation experiment, we qualitatively compared navigation sentences generated by objects selected from the driver's estimated gaze with those selected on the basis of class probabilities. The results showed that utilizing a gaze estimation model trained with DGD allows us to select objects that are easily visible to the driver and avoid the selection of objects that provide incorrect instructions.

Our method successfully achieved HℓG in a simple intersection scene. However, we recognize its potential for further development. By expanding the number of classes, defining additional templates, and enhancing the performance of the gaze estimation model, we can extend our method to handle more complex intersection scenes and produce human-like navigation instructions. Furthermore, this time we use the names of specific objects, but they could be simpler. This advancement has the potential to address existing challenges in car navigation systems and contribute to the advancement of HℓG as a whole.

References

[1] A. K, Kandangath, and X. Tu, "Humanized navigation instructions for mapping applications," April 23 2015. US Patent application. 14/061,208.

[2] B. Wang, Q. Stafford-Fraser, P. Robinson, E. Dias, and L. Skrypchuk, "Landmarks based human-like guidance for driving navigation in an urban environment," IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 1–6, 2017.

[3] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," Proceedings of Machine Learning Research (PMLR), vol. 37, pp. 2048–2057, 2015.

[6] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Neural baby talk," Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7219–7228, 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," In Advances in Neural Information Processing Systems (NIPS), vol. 30, 2017.

[8] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," Neural Information Processing Systems, no. 999, pp. 11 137–11 147, 2019.

[9] V.-Q. Nguyen, M. Suganuma, and T. Okatani, "Grit: Faster and better image captioning transformer using dual visual features," European Conference on Computer Vision (ECCV), pp. 167–184, 2022.

[10] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2798–2805, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[13] M. Kummerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low- and high-level contributions to fixation prediction," International Conference on Computer Vision (ICCV), 2017.

[14] S. Jia and N. D. B. Bruce, "Eml-net:an expandable multi-layer network for saliency prediction," Image and Vision Computing, vol. 95, 2020.

[15] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," IEEE Transactions on Multimedia, vol. 22, no. 8, pp. 2163–2176, 2019.

[16] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 636–644, 2017.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 2015.

[19] I. Kasahara, S. Stent, , and H. S. Park, "Look both ways: Self-supervising driver gaze estimation and road scene saliency," European Conference on Computer Vision (ECCV), pp. 126–142, 2022.

[20] S. Sourabh Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," IEEE Transactions on Intelligent Vehicles, vol. 3, pp. 254–265, 2018.

[21] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," Proceedings of Machine Learning Research (PMLR), vol. 78, pp. 1–16, 2017.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2021.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," Conference on computer vision and pattern recognition, pp. 248–255, 2009.

[24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[25] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," International Conference on Computer Vision (ICCV), pp. 1314–1324, 2019.

[26] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," Proceedings of Machine Learning Research (PMLR), 2021.

[27] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," Transactions on Machine Learning Research (TMLR), 2023.