Pedestrian's Gaze Object Detection in Traffic Scene

Hiroto Murakami[1] [a], Jialei Chen[1], Daisuke Deguchi[1],
Takatsugu Hirayama[2,1], Yasutomo Kawanishi[3,1] and Hiroshi Murase[1]

[1]*Graduate School of Informatics, Nagoya University, Nagoya, Japan*
[2]*Faculty of Environmental Science, University of Human Environments, Okazaki, Japan*
[3]*Multimodal Data Recognition Research Team, Guardian Robot Project, Riken, Kyoto, Japan*
hiroto.murakami@nagoya-u.jp

Keywords: Pedestrian's Gaze Object Detection, Object Detection, Gaze Estimation, Traffic Scene, Dataset.

Abstract: In this paper, we present a new task of detecting an object that a target pedestrian is gazing at in a traffic scene called PEdestrian's Gaze Object (PEGO). We argue that the detection of gaze object can provide important information for pedestrian's behavior prediction and can contribute to the realization of automated vehicles. For this task, we construct a dataset of in-vehicle camera images with annotations of the objects that pedestrians are gazing at. Also, we propose a Transformer-based method called PEGO Transformer to solve the PEGO detection task. The PEGO Transformer directly performs gaze object detection with the utilization of whole-body features without a high-resolution head image and a gaze heatmap which the traditional methods rely on. Experimental results showed that the proposed method could estimate pedestrian's gaze object accurately even if various objects exist in the scene.

1 INTRODUCTION

Detection of PEdestrian's Gaze Object (PEGO) aims to detect an object that a pedestrian is gazing at in a traffic scene. This is an important task for computers to predict the future behavior of a pedestrian in a traffic scene. For example, as shown in Fig. 1, a pedestrian gazes at an oncoming car and will probably wait until the car passes without jumping out into the roadway. Thus, the detection result of the gaze object is an important clue that reveals what behavior the person intends to take in the future.

For the gaze detection task, several datasets have been released. Recasens et al. (Recasens et al., 2015) have released a GazeFollow dataset for gaze detection in our daily life. This pioneering work demonstrates the importance of gaze detection tasks in person behavior prediction. Tomas et al. (Tomas et al., 2021) have released a Gaze On Objects (GOO) dataset that aims to find products that customers are gazing at in the retail store scene.

Several gaze detection models have been developed using these datasets. Recasens et al. (Recasens et al., 2015) have proposed a model to detect line of sight using the GazeFollow dataset. Wang et

Figure 1: A pedestrian is gazing at an oncoming vehicle.

al. (Wang et al., 2022) used the GOO dataset and have proposed a GaTector that detects human's gazing objects.

In a traffic scene, pedestrian's gaze detection is also essential because it contributes to determining automated driving behavior and implementing technologies that alert drivers. Belkada et al. (Belkada et al., 2021) and Hata et al. (Hata et al., 2022) have proposed a method for detecting "eye contact", which indicates whether a pedestrian is gazing at the in-vehicle camera. However, they have not been studied to recognize which objects pedestrians are gazing at if they are not gazing at the in-vehicle camera.

A dataset plays an important role in achieving pedestrian's gaze object detection in traffic scenes. Various datasets have been released for traffic-scene

[a] https://orcid.org/0009-0008-6571-4721

understanding (Caesar et al., 2020; Sun et al., 2020; Rasouli et al., 2019; Cordts et al., 2016; Geiger et al., 2013). However, to the best of our knowledge, there is no dataset consisting of annotations on pedestrian's gaze objects in traffic scenes. Since existing datasets and methods focus on our daily lives and retail store scenes, they are different domains and cannot be used for PEGO detection in traffic scenes.

Therefore, this paper tackles the tasks to detect the pedestrian's gaze object: construct a new dataset and propose a new method. In this dataset, we manually annotate each pedestrian in an in-vehicle camera image with the pedestrian's gaze object. In addition, we propose a method termed PEGO Transformer for detecting the pedestrian's gaze object using this dataset. The PEGO Transformer consists of four modules: a backbone to extract features from the input images, a Deformable Transformer to capture the features corresponding to objects, a Projection Layer to utilize the features to produce the result for the final prediction, and a Label Generator to generate the label index for training the model via loading the dataset.

Contributions of this paper are as follows.

1. This paper proposes a novel PEGO Transformer that can estimate a gazing object of each pedestrian in a traffic scene. The proposed method is capable of estimation even without the high-resolution head images or the gaze heatmap required by conventional methods. The PEGO Transformer is trained to capture the relationship between detected objects and pedestrians so that the likelihood of gaze object for each pedestrian becomes high.

2. This paper proposes a novel task of PEGO detection that estimates gaze object of each pedestrian in the traffic scene. For this task, we construct a new dataset by extending the widely used traffic scene dataset.

2 RELATED WORK

2.1 Human's Gaze Object Detection

Recasens et al. (Recasens et al., 2015) constructed a gaze detection dataset and proposed a method called GazeFollow to estimate human gaze. The aim of their dataset is to estimate the direction of the gaze, whereas our study aims at detecting the gaze object.

Wang et al. (Wang et al., 2022) and Tu et al. (Tu et al., 2022; Tu et al., 2023) proposed methods for detecting the gaze object of a target human at the object level. Wang et al. proposed GaTector, which estimates the products customers are gazing at in a store scene. The gaze heatmap is estimated using a high-resolution head image of the target person, and objects that overlap with the estimated gaze heatmap are considered the gaze objects. Human-Gaze-Target Detection with Transformer (HGTTR) (Tu et al., 2022) and Gaze following detection Transformer (GTR) (Tu et al., 2023) proposed by Tu et al. detect human gaze object in more general scenes. As with GaTector, Tu's methods are processed in the head detection branch and the gaze heatmap detection branch, after which the gaze object is estimated. However, these methods rely on high-resolution head images. Most pedestrians captured by in-vehicle cameras are smaller than those captured in the other scenes due to their distance and lower image resolution. This makes it difficult to extract the head features required for estimating the region of attention by these methods. In addition, since the gaze target is selected based on the estimated gaze heatmap, the selection accuracy is highly dependent on the performance of the gaze heatmap estimator. As a result, our method can estimate the gaze object without a high-resolution head image by using the pedestrian's whole body features. It also estimates the gaze object directly without using a gaze heatmap, that is a performance bottleneck of the previous methods.

2.2 Pedestrian's Gaze Target Detection

Belkada et al. (Belkada et al., 2021) and Hata et al. (Hata et al., 2022) worked on pedestrian eye-contact detection. They use skeletal information to detect whether a pedestrian is gazing at the in-vehicle camera because pedestrians captured by in-vehicle cameras are often small and blurred, and thus existing eye gaze detection methods cannot be applied directly. They also constructed new datasets that can handle the eye contact detection task by extending an existing traffic scene dataset. The task addressed in our study is similar to that addressed by Hata et al. in terms of focusing on pedestrians captured by an in-vehicle camera. However, their method cannot recognize the pedestrian's gaze object without any eye-contact.

2.3 Dataset Containing Pedestrians

Datasets recorded in real traffic scenes are beneficial for automated driving tasks. Caesar et al. (Caesar et al., 2020) have released nuScenes and nuImages annotated with bounding boxes and object class labels for object detection. In these datasets, bounding
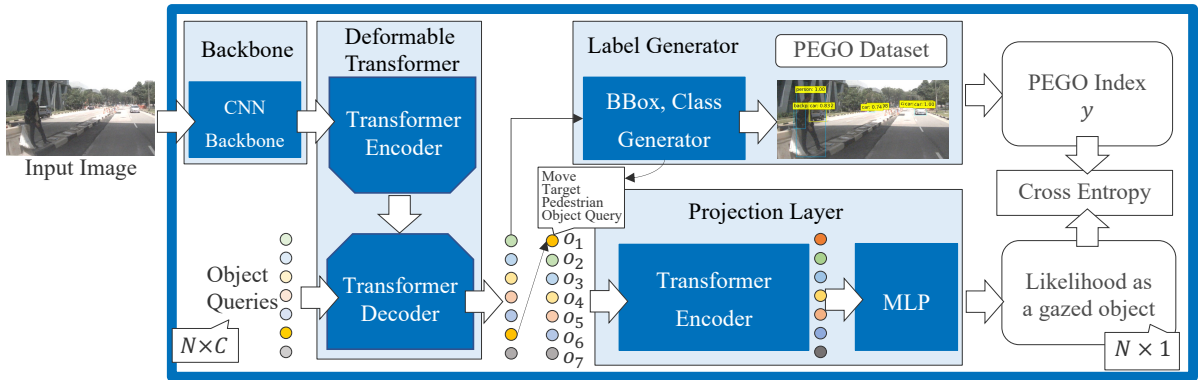
Figure 2: Overview of the PEGO Transformer.

boxes and 23 classes of object labels are annotated for vehicles, bicycles, and pedestrians captured by in-vehicle cameras. However, the state of each pedestrian, such as the gaze direction and the gaze object, is not annotated.

Sun et al. (Sun et al., 2020) have released the Waymo Open Dataset (Waymo), which is annotated with object bounding boxes and class labels in the same way as nuScenes and nuImages. However, the state of pedestrians is neither annotated.

On the other hand, the Pedestrian Intention Estimation dataset (PIE dataset) (Rasouli et al., 2019) is the dataset constructed for pedestrians' behavior prediction. In this dataset, 1,842 pedestrians captured by in-vehicle cameras are annotated with information such as ID, bounding box, whether they are likely to cross the road, and whether they are gazing in the camera direction. However, the pedestrian states necessary for the PEGO detection task, such as the gaze direction and object, are not annotated.

3 PEDESTRIAN'S GAZE OBJECT DETECTION

In this section, we propose the PEGO Transformer, which detects pedestrians' gaze object in the image. Unlike conventional gaze object detection methods, our method does not rely on a high-resolution head image, but uses features from a full-body image for PEGO estimation. Also, instead of relying on a gaze heatmap, the PEGO Transformer is trained to capture the relationship between detected objects and pedestrians so that the likelihood of gaze object for each pedestrian becomes high. The architecture of the PEGO Transformer is shown in Fig. 2. The architecture consists of four modules: a backbone to extract features from the input images (CNN backbone), a Deformable Transformer (Zhu et al., 2021) to refine
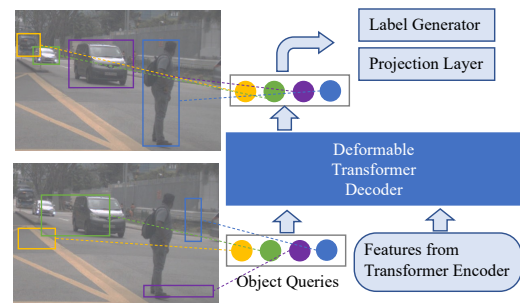

Figure 3: Object queries.

the features from backbone (transformer encoder and decoder), a Projection Layer to utilize the features to produce the result for the final prediction, and a Label Generator to generate the label index for training the model via loading the dataset. We introduce each of these modules in the following section.

3.1 Architecture of PEGO Transformer

Backbone. The backbone, consisting of a CNN, aims to produce features with high-level semantics for input to the Deformable Transformer. Given an input image $x \in \mathbb{R}^{C \times H \times W}$, x is first fed into the CNN backbone (*e.g.*, ResNet (He et al., 2016)), to provide features with high-level semantics.

Deformable Transformer. The Deformable Transformer, consisting of a Deformable Transformer Encoder and Deformable Transformer Decoder (Zhu et al., 2021), aims to produce features that corresponds with each object. The features from backbone are then flattened and combined with positional encoding for the deformable transformer encoder. In the deformable transformer encoder, which benefits from the deformable self-attention module, the features interact with each other to enhance the output.

As shown in Fig. 3, the Deformable Transformer Decoder takes the output of the feature extractor as input and associates the features corresponding to each

object to an object query ($o \in \mathbb{R}^C$) (Carion et al., 2020) using the following procedure. First, the $O = \{o_1, o_2, ..., o_N\}$ are initialized with random values and input to the deformable transformer decoder. The input O are associated with the output of the Transformer Encoder and the features corresponding to the object. And then the deformable self-attention module captures the relationship between each $o \in O$. As a result, each $o \in O$ becomes a feature that corresponds one-to-one with the object.

Projection Layer. The Projection Layer (Vaswani et al., 2017) consists of a Transformer Encoder and an MLP layer. The Transformer Encoder captures the gazing correspondences between pedestrians and objects with O as input. The MLP aims to produce the confidence scores indicating if an object is gazed.

Label Generator. The Label Generator outputs the label index of the pedestrian's gaze object for training the model. When a pedestrian is gazing at o_y corresponding to an object, then the output label index is y.

To generate the label index, we first estimate the bounding box and class probability of the object from each $o \in O$. Next, we select o_m that has the highest probability for the pedestrian class and whose value exceeds the threshold δ. Then, we store m which is the index of the pedestrian to $\mathcal{M} = \{m_1, m_2, ...\}$. Finally, the gaze object of the pedestrian corresponding to $m \in \mathcal{M}$ is obtained from the dataset, and outputs the index, y, when o_y corresponds to the gaze object. The index $m \in \mathcal{M}$ of the pedestrian object query obtained from the output of the label generator is also used in the projection layer to sort the target pedestrian object queries to the top.

3.2 Loss Function

The cross-entropy is used as the loss function. The cross-entropy loss is calculated from the softmax of the likelihood of each object as a gaze object by the projection layer and the label of the gazed object output by the label generator.

3.3 Inference of Pedestrian's Gaze Object

During the inference procedure, to detect the gaze object, the in-vehicle camera image is first input to the PEGO Transformer to obtain the likelihood of each object as a gaze object. The object with the highest likelihood is selected as the gaze object.

Table 1: Number of pedestrians and images in the constructed dataset.

Source Dataset	Pedestrians	Images
nuScenes	292	218
nuImages	1,240	870
Waymo	1,193	672
Total	**2,725**	**1,760**

4 PEDESTRIAN'S GAZE OBJECT DATASET

To verify the performance of the PEGO Transformer, we construct a PEGO Dataset annotate with the pedestrian's gaze object. In contrast to existing studies, we annotate pedestrians' gaze points in each image. Our dataset contains annotations of the target pedestrian's ID, bounding box coordinates, gaze point coordinates, and pedestrian status. If the gaze object cannot be identified, we annotate the point at which a pedestrian is gazing. In addition, when it is difficult to judge the point being gazed at, such as in the case of eye contact or backward facing, we record these situations as an additional annotation in the dataset. Three annotators annotate the same image to maintain the quality of the annotation. Details of the dataset are described in the following sections.

4.1 Image Details

This dataset was constructed based on the existing datasets: nuScenes, nuImages and Waymo (Caesar et al., 2020; Sun et al., 2020). These are large open datasets containing images captured by in-vehicle cameras and are annotated with the object's bounding box and its class label, as described in Section 2.3.

In our dataset, only images satisfying the following conditions were collected:

1. The overlap between a pedestrian and other object is less than 25 %.

2. The height of the pedestrian bounding box is 200 pixels or more.

3. The entire pedestrian bounding box is present in the in-vehicle camera image.

4. Target pedestrian and the gaze object appear in the same image.

5. The annotation target frames are selected by every 5 seconds in nuScenes and Waymo.

Consequently, a total of 2,725 pedestrians (1,760 images) were selected for annotation. The detail of the dataset is shown in Table 1.
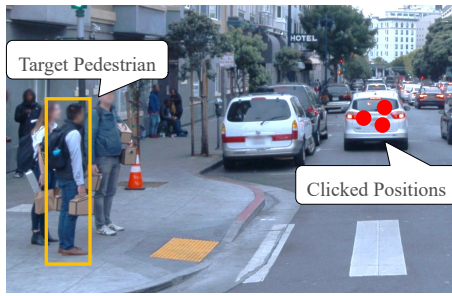
Figure 4: Example of annotation.

(a) No object   (b) Outside the image

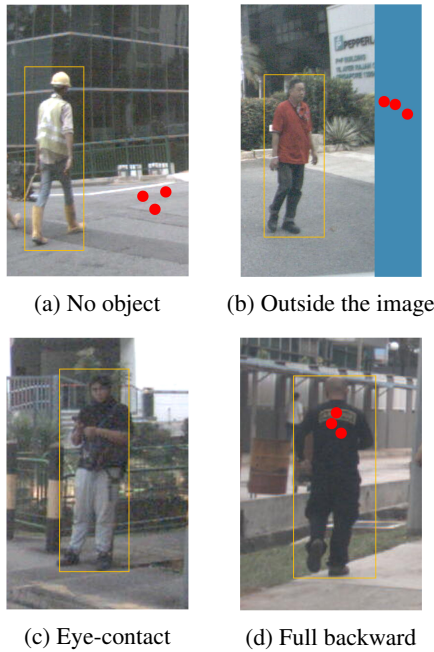(c) Eye-contact   (d) Full backward

Figure 5: Situations in which annotators are unable to select an object that the pedestrian is gazing at.

4.2 Annotations

Three annotators annotated 2,725 pedestrian's gaze objects in the dataset. As shown in Fig. 4, the annotators clicked on the gaze object of each pedestrian using a specialized annotation tool for this task.

In some cases, the annotators could not determine the gaze object due to several reasons, such as no gaze object in the scene, the gaze object outside the image, eye-contact, and full backward posture. Such pedestrians cannot be annotated directly by the above annotation steps. For such pedestrians, the annotators annotated the special labels as follows:

No Object

As shown in Fig. 5(a), the "no object" label is annotated for a pedestrian who is not gazing at any object. In this case, the annotator clicked on the area at which the pedestrian was gazing.

Table 2: Annotation results.

Gazing at	Pedestrians
Object	1,234
No object	450
Outside the image	370
Eye-contact	256
Full backward	20
Others	395

Outside the Image

As shown in Fig. 5(b), the "outside the image" label is annotated for a pedestrian whose gaze object outside the image. We placed a small clickable area around the image in the annotation tool, and the annotator clicked within this area while maintaining the pedestrian's viewing direction.

Eye Contact

As shown in Fig. 5(c), the "eye-contact" label is annotated for a pedestrian who was gazing at the in-vehicle camera.

Full Backward

As shown in Fig. 5(d), the "full backward" label is annotated for a backward-facing pedestrian, because it is difficult to determine the gaze object in the scene.

4.3 Annotation Results

Table 2 and Fig. 6 show the annotation results of the pedestrian's gaze objects in our dataset. As previously mentioned, 2,725 pedestrians were annotated. The target pedestrians are indicated by the yellow boxes, the red dots are the points clicked by the annotators, and the objects indicated by the red boxes are PEGOs in Fig. 6. Each object recorded how many annotators selected it as a PEGO. This allows the ground truth of the gaze object to be changed according to the PEGO detection task. In contrast to existing datasets for gaze estimation, the size of the pedestrian relative to the image size is small, and the pedestrian and the target object are far apart.

The following annotations were included in the dataset.

- Target pedestrian's ID
- Target pedestrian's bounding box
- Gaze point coordinates
- Eye-contact or not
- Full backward or not
- Bounding box of the PEGO (only for a pedestrian who gazes at an object)
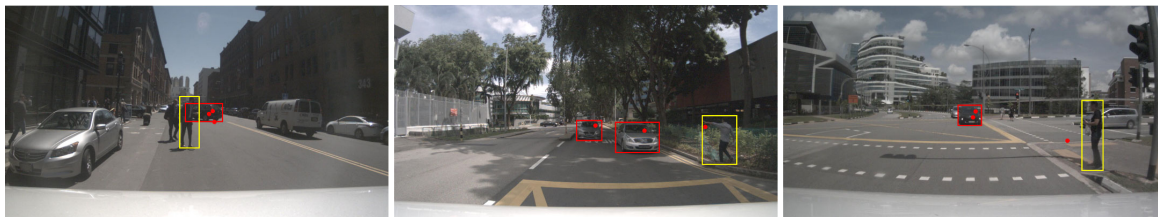- Category of the PEGO (only for a pedestrian who gazes at an object)

Figure 6: Annotation results: The target pedestrian in each image is indicated by yellow box and the PEGOs are indicated by red boxes. Gaze points in the annotation results are indicated by red dots.

5 EXPERIMENT

A trained PEGO Transformer was used to detect the pedestrian's gaze object. In this section, we present the experimental conditions and the results.

5.1 Implementation Details

We trained the PEGO Transformer on the dataset explained in section 4. Pedestrians gazing at the object were used for training. We performed five-fold-cross-validation on the dataset. Each fold contains 47, 44, 43, 44, 48 scene images and 92, 91, 89, 112, 130 pedestrians, respectively. In training step, horizontal flips were applied to each image as data augmentation.

In the training procedure, only the parameters of the projection layer were updated. The feature extractor and the deformable transformer were initialized with the pre-trained weights of the Deformable DETR (Zhu et al., 2021). We used the top 40 object queries ($N = 40$) whose highest-class probability was higher than the threshold. The threshold δ was set to 0.3.

5.2 Comparative Methods

To investigate the effectiveness of PEGO Transformer, we used two comparative methods to estimate the gaze object of pedestrians. The first model was a line-of-sight prediction based on GazeFollow (Recasens et al., 2015). We created an MLP model to estimate the pedestrian's line of sight and trained it on the constructed PEGO dataset. The distance from the estimated line of sight to each of the candidate objects was calculated, and the object closest to the estimated line of sight was selected as the gaze object.

The second model was GaTector (Wang et al., 2022). The model was pre-trained on the GOO dataset (Tomas et al., 2021) and fine-tuned on the constructed PEGO dataset. In this comparison experiment, the energy aggregation loss of GaTector was

calculated for all candidate objects in the image using the estimated gaze heatmap, and the object with the smallest loss was selected as the gaze object. The energy aggregation loss is the ratio of the average of the estimated gaze heatmap over the entire image to the average of the estimated heatmap over the bounding box of the object.

5.3 Results and Discussion

Table 3 reports the accuracy of the PEGO detection. We evaluated whether the detected gaze object with the highest score for the target pedestrian was the same as the ground truth of the gaze object in our dataset, which we refer to as the Top1 accuracy. In addition, we evaluated whether the correct answer could be included within the 2nd, 3rd, 4th, and 5th highest objects, referred to as Top2, Top3, Top4, and Top5 accuracy, respectively.

The PEGO Transformer is able to estimate the gaze object of pedestrians with higher accuracy than chance rate and comparison methods. Figures 7(a), (b), and (c) show examples of successful PEGO detection. From these results, the PEGO Transformer succeeded in estimating the gaze object using full-body image features. In contrast, the estimation result in Fig. 7(d) was the opposite direction from her as the gaze object.

As seen in Table 3, the proposed PEGO Transformer outperformed the comparative method (GaTector) that explicitly uses head images. GaTector requires high-resolution head images to estimate PEGO, but it is difficult to obtain such images of target pedestrians in traffic scenes because their distances become high. Therefore, the performance of GaTector was lower than expected.

Next, the line of sight based method is difficult to determine gaze object when multiple objects exist close to the line of sight. On the other hand, as seen in Fig. 7(c), the proposed PEGO Transformer can correctly estimate the gaze object even if a pedestrian stands very close to the target pedestrian. However, the proposed PEGO Transformer does not take into

Table 3: Pedestrian's Gaze Object (PEGO) detection accuracy.

	Top1(%)	Top2(%)	Top3(%)	Top4(%)	Top5(%)
Random guess	3.44	6.78	10.0	13.1	16.2
Comparative (Line of sight)	29.8	51.4	56.8	63.9	65.3
Comparative (GaTector)	19.5	22.3	24.6	29.3	32.0
Proposed (PEGO Transformer)	**47.2**	**70.8**	**83.3**	**93.8**	**97.9**

(a) True: She gazes an approaching car.

(b) True: She gazes an approaching car.

(c) True: He gazes the car in the back.

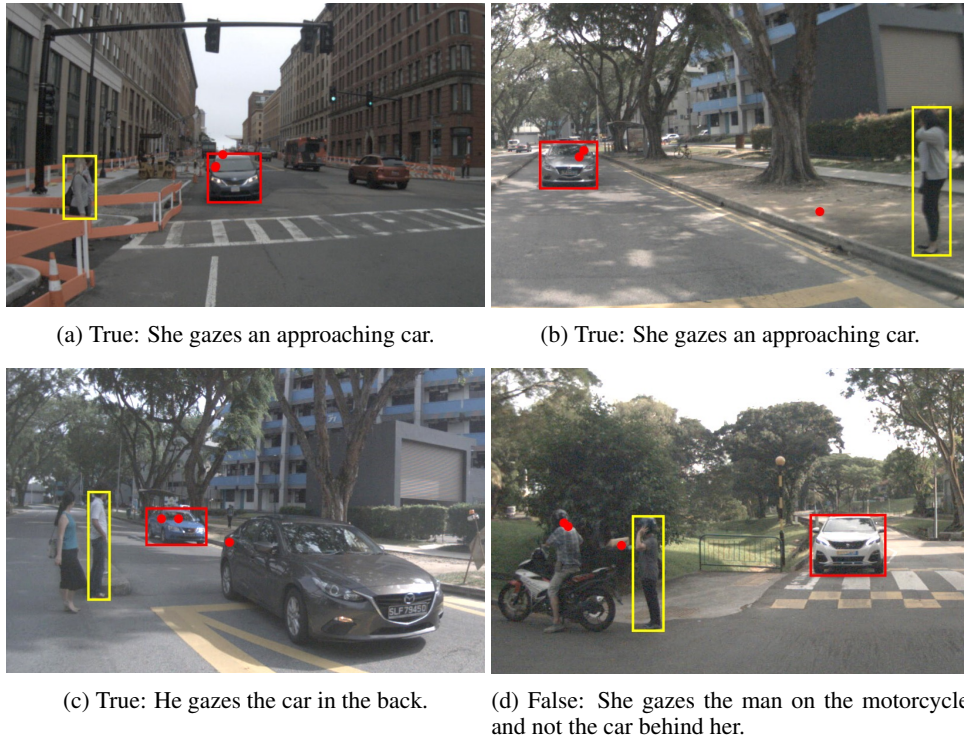(d) False: She gazes the man on the motorcycle and not the car behind her.

Figure 7: Examples of detection: The target pedestrian is indicated by the yellow box and the estimated PEGO by the red box. Gaze points in the annotation results are indicated by red dots.

account the pedestrian's pose, which can make it difficult to determine the gaze object, as seen in Fig. 7(d).

As analyzed by Wang et al. the bottleneck of GaTector is that the results of the gaze heatmap estimation affect the gaze object prediction (Wang et al., 2022). On the other hand, the PEGO Transformer can directly detect the gaze object without relying on the gaze heatmap. Therefore, the PEGO Transformer performed well in scenes where gaze heatmap estimation was difficult.

6 CONCLUSIONS

In this paper, we present a new task of detecting an object that a target pedestrian is gazing at in a traffic scene. Our proposed PEGO Transformer can estimate a pedestrian's gaze object without the high-resolution

head image and the gaze heatmap used in conventional gaze detection methods. Unlike existing gaze detection datasets considering human daily lives, our dataset focuses on traffic scenes. This method and dataset can provide important information for behavior prediction and contribute to the realization of automated vehicles.

The PEGO dataset proposed in this paper consists of gazed objects of pedestrians that are annotated from a third-person view. This annotation scheme can be easily applied to existing large data sets, but the annotations may differ from the true objects gazed at by the pedestrians. Thus, future work will include evaluation of the PEGO transformer in controlled experiments in which pedestrians gaze at predefined targets.

The previous study by Hata et al. (Hata et al., 2022) proved that skeleton information is effective in estimating whether a pedestrian is gazing at an in-

vehicle camera. Therefore, it is expected that the accuracy will be further improved by taking skeletal information into account in the PEGO Transformer. We plan to extend the dataset to improve the accuracy of the proposed method.

ACKNOWLEDGMENT

REFERENCES

Belkada, Y., Bertoni, L., Caristan, R., Mordan, T., and Alahi, A. (2021). Do pedestrians pay attention? eye contact detection in the wild.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11618–11628.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-End object detection with transformers. In *In Proceedings of the European conference on computer vision*, pages 213–229. Springer.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, page 1231–1237.

Hata, R., Deguchi, D., Hirayama, T., Kawanishi, Y., and Murase, H. (2022). Detection of distant eye-contact using spatio-temporal pedestrian skeletons. In *Proceedings of the IEEE 25th International Conference on Intelligent Transportation Systems*, pages 2730–2737.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Rasouli, A., Kotseruba, I., Kunic, T., and Tsotsos, J. (2019). PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, pages 6261–6270.

Recasens, A., Khosla, A., Vondrick, C., and Torralba, A. (2015). Where are they looking? In *Proceedings of the Advances in Neural Information Processing Systems*, volume 28, pages 199–207.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D. (2020). Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2443–2451.

Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R., and Guinto, R. (2021). GOO: A dataset for gaze object prediction in retail environments. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 3119–3127.

Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., and Shen, W. (2022). End-to-End Human-Gaze-Target Detection with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2192–2200.

Tu, D., Shen, W., Sun, W., Min, X., Zhai, G., and Chen, C. (2023). Un-gaze: a unified transformer for joint gaze-location and gaze-object detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is All you Need. In *Proceedings of the 2017 Advances in Neural Information Processing Systems*, volume 30.

Wang, B., Hu, T., Li, B., Chen, X., and Zhang, Z. (2022). GaTector: A unified framework for gaze object prediction. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19588–19597.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). Deformable DETR: Deformable transformers for end-to-end object detection. In *Proceedings of the 9th International Conference on Learning Representations*.