Appendices for Modeling Drivers’ Situational Awareness from Eye Gaze for Driving Assistance

Anonymous Author(s)
Affiliation
Address
email

A Additional Related Work

A.1 Situational Awareness: definitions from aviation to driving

First popularized by Mica Endsley’s work in aviation, pilots’ SA was defined as “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” [1]. According to Endsley, SA reflects the extent to which the operator knows what is going on in their environment and is the product of mental processes including attention, perception, memory, and expectation [2]. This definition laid out three levels of SA: (1) perception (of situational elements) , (2) comprehension (of their semantics), and (3) projection (of their futures states). In the original aviation context, these elements comprised instruments and instrument panels that pilots needed to maintain SA over in order to perform the aviation task safely and successfully. However, in the driving context these scene elements not only comprise similar in-vehicle instruments such as the speedometer and rear-view mirrors, but also outside-the-vehicle elements such as other vehicles, bicycles, pedestrians etc. For tracking with respect to pilot/driver eye gaze, a functionally challenging difference among these elements is that the driving elements constantly change position relative to the vehicle while the aviation instruments are fixed and their locations are known. This difference makes is difficult to apply techniques (for grounding, evaluation etc.) from aviation directly to the driving case.

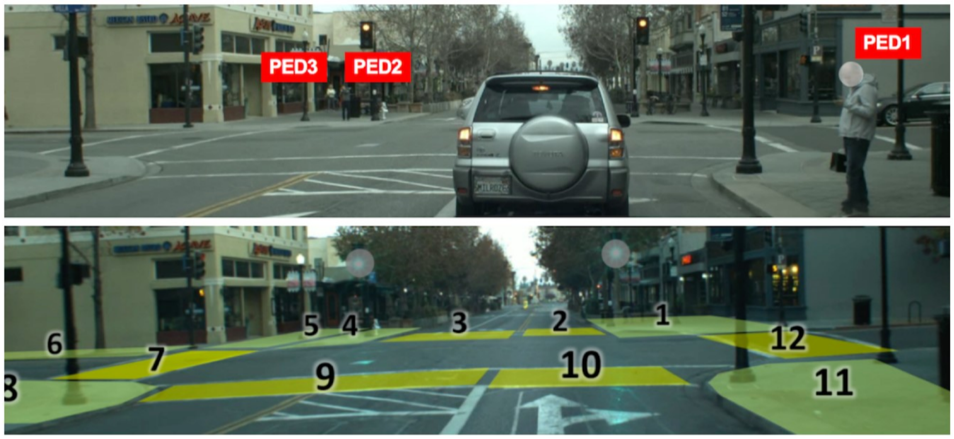
A.2 Situational Awareness labeling methods

At a high level, situation awareness (SA) grounding methods can be classified into direct (e.g. queries about objects for which SA is estimated) and indirect (SA inferred from secondary task measures such as response time to probes). As we discuss these, we will comment on the suitability of these techniques to generate per-object labels for learning a gaze-based per-object SA model.

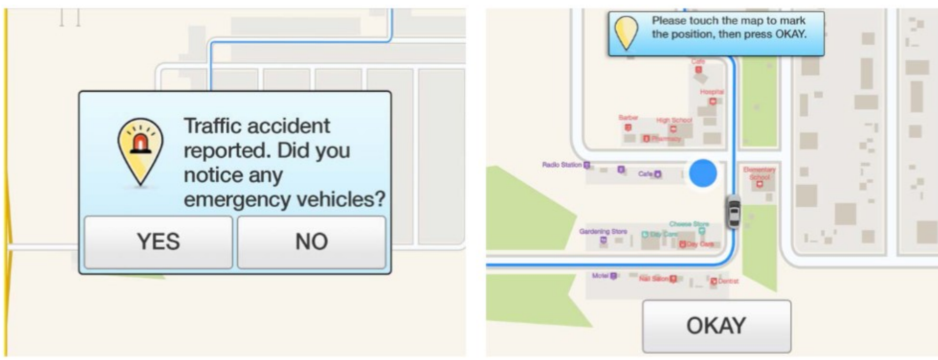
SA Labeling Method	Capture Awareness Transition	Dense Object Labels	Doesn’t Affect Natural Gaze Behaviour
SAGAT [5]	×	✓	✓
DAZE [4]	✓	×	×
SPAM [6]	✓	×	×
Ours	✓	✓	✓

Table 1: Our SA labeling protocol allows us to capture the transition in the driver’s awareness of objects in the scene, allows labels for all objects in the scene without affecting the natural gaze behaviour of the driver.

Submitted to the 8th Conference on Robot Learning (CoRL 2024). Do not distribute.



(a) SAGAT freezes simulations or videos being watched (top) and then asks participants the location of traffic elements (bottom). Image from [3].



(b) DAZE does not require pauses. It asks participants if they noticed particular types of traffic elements and to mark their locations on an overhead GPS map. Image from [4].

Figure 1: Examples of SA labeling methods used in previous work. These methods produce intermittent labels (SAGAT/DAZE) or sparse ones (DAZE —not every object is labeled).

A.2.1 Direct methods

Within direct methods, we may classify grounding techniques into objective or subjective based on whether the probes involve questions about directly measureable quantities (e.g. number of red vehicles around you) or self-rated ones (e.g. perceived task load). We will first discuss objective measures. Perhaps the most well known and used direct objective method of Situational Awareness grounding is the Situation Awareness Global Assessment Technique (SAGAT) [7]. The SAGAT involves operators performing a simulated version of a real task such as driving. Intermittently, the simulation is paused (the screen can be blanked or only the background is presented) and the operators are asked several questions about the situation right before the pause. Accuracy of responses to these questions determines the operators' SA. SAGAT was first designed for aviation but has been adapted to driving [3]. Despite its popularity, SAGAT has its limitations mainly associated with the mandatory simulation pauses required. There are cognitive process modifications to the normal task because of removal from the task during the probe as well as intermittent task resumption deviations [8]. For generating ground truth data for per-object SA, we also have some issues. One, we only get SA labels per queried object at the time of the probe —SAGAT probes do not give us the starting point of the operators' SA for each queried object. Second, SAGAT querying requires pauses hence limiting the number of labels per drive that could be collected while maintaining the flow of simulation.

Another direct objective measure that mitigates some of these issues is Daze [4] which uses real-time in situ questions that resemble queries drivers are already familiar with (such as traffic queries from apps like Waze). In particular, shortly after an on-road event such as an accident has passed, it raises an alert asking a question such as “Traffic accident reported. Did you notice any emergency vehicles?”. While this method avoids pausing the simulation (an indeed can also be used for on-road driving), it does not provide dense, per-object labels in the way we require. Additionally, answering the query involves looking away from the driving scene and at a tablet or screen which undesirably modifies gaze behavior.

In conjunction with objective methods, subjective measurements can be useful. For example, operators’ perceived estimate of their own SA may important in determining their actions or interactions with an SA enhancing system. Here, we will only discuss the most commonly used subjective measure: Situational Awareness Rating Technique (SART). SART is administered as a 14-part post-hoc questionnaire in which, operators rate on a series of bipolar scales the degree to which they perceive (1) a demand on their resources, (2) supply of operator resources and (3) understanding of the situation. These are combined to provide an overall SART score [9]. However, there are limitations to SART as a measure of the operators’ SA. For example, consider unknowingly unknown scene elements: operators cannot rate their SA on all scene elements if they didn’t know they missed some. Other factors are the influence of performance on SART, as well as confounding with workload [10].

A.2.2 Indirect methods

Within indirect SA grounding techniques, the most widely accepted protocol is the Situation Present Awareness Method (SPAM) [6]. SPAM involves a real-time probe (usually a verbal query about the past, present, and future aspects of the situation) while the operator is performing their primary task. While direct measures such as response accuracy are collected, SPAM importantly also uses response times as an index of how readily this information is available. For our requirements, verbal queries have the same label sparsity issue as Daze as well as requiring manual post-processing to get machine readable annotations from verbal responses.

A.2.3 Physiological methods

For the sake of completeness we must mention the use of physiological methods in the literature to measure operator SA. These signals have the benefit of being continuous variables rather than isolated or posthoc probes mentioned above. These methods have employed physiological signals such as EEG [11], respiratory rate [12], and heart rate [13] to measure SA. Of these methods, EEG has the most predictive power, while respiratory measures were found to have a negative correlation with SA [14].

The most commonly used physiological technique was based on eye tracking. This included signals as blink rates, pupil dilation, but also behavioral characteristics such as fixation rates, dwell times, and saccade frequency to measure SA [14].

However, physiological methods are noisy, show small correlations with SA, and only provide an overall impression of SA rather than per-object SA. The most promising physiological modality was eye gaze, with eye tracking based features forming the best performing predictors of SA. For a full treatment of this topic we refer the reader to Zhang et al. [14].

B Situational Awareness Data Collection

We use DReyeVR [15] as the VR-driving simulator. DReyeVR extends the Carla [16] simulator to add virtual reality integration, a first-person maneuverable ego-vehicle, eye tracking support, and several immersion enhancements such as mirrors and sounds. Our physical setup includes a HTC Vive Pro Eye as the head-mounted VR device, which has built-in eye tracking, and an available eye tracking SDK. For our driving hardware we use a Logitech G29 wheel and pedals kit. For driving routes, we use custom routes from several virutal towns shipped with CARLA. Furthermore, we

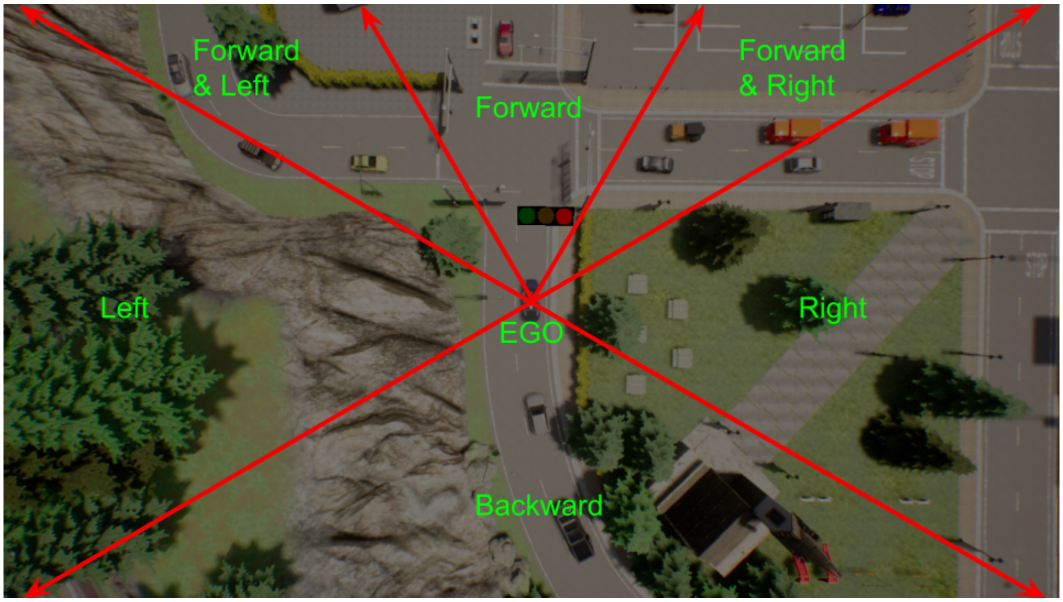


Figure 2: Sectors corresponding to the directions of the button presses. The objects in each sector are target objects for button presses corresponding to the direction of the sector.

control the traffic in the simulation such that only a single vehicle or two-wheeler enters the FoV of the driver from a single direction at an intersection. If multiple objects enter the driver's FoV from the same direction at the same time, even if the user presses the corresponding directional buttons multiple times, we use manual post-hoc annotation to resolve ambiguities for button press assignment to objects.

B.1 Instructions provided to participants:

The following prompt was read to participants before they underwent the first trial route. *"Drive safely while following signs to the goal destination. Your main objective is to arrive at the destination as quickly as possible while driving safely. While doing so, you will also perform a secondary task by pushing buttons to indicate which vehicles, pedestrians or two-wheelers (collectively, traffic objects) you have perceived in the environment around you. Anytime you see a new vehicle please press one of the four arrow key on the left side of your steering corresponding to the direction in which they first appeared in your field of view. Similarly, for pedestrians and two-wheelers use the 4 buttons on the right. For each new traffic object you should only press the button once."*

B.2 SA label inference from button presses

Our SA protocol as described in Sec 3 of the paper, allows users to indicate their awareness of objects in the scene using directional button presses. The direction of the button corresponds to the direction of the object. Additionally, there are two sets of directional buttons for the users to choose from. One set corresponds to vehicles and the other set corresponds to pedestrians (+ two-wheelers). For example, when a user first becomes aware of a pedestrian on their left, they would press the left directional button from the button set corresponding to pedestrians.

Our protocol provides us with button clicks, to convert these into awareness labels for object we need to associate button clicks with objects in the scene. We rely on the direction and the set of the button press to associate button presses with objects. We divide the entire scene into 4 sectors corresponding to the 4 directional buttons. (Fig 2). The top sector corresponds to the area between +30 and -30 degree from the ego vehicle. The left sector corresponds to the area between -60 and -120 degree, the right sector corresponds to the area between +60 and +120 degree. The back sector lies between -120 and +120 degrees. The sector between +30 and +60 is considered both forward and right, similarly the sector between -30 and -60 is considered both forward and left.

We keep a track of all the objects that enter each sector, and associate objects with the button clicks pertaining to each sector. The object in each sector, which has not been associated with any button clicks can be associated with a new button click. Objects are considered aware once they are associated with a button click, however once they re-enter of the field-of-view of the driver after leaving it for a certain amount of time, they are again considered unaware and can be associated with button clicks again.

We control the traffic to ensure that there are only a single object of each type (vehicle, pedestrian) in each sector. However, to add randomness we also add a very small number of randomly spawned objects in the scene. Due to this, in certain situations participants' button press inputs can be ambiguous relative to the traffic scene. One common scenario involved multiple potential target objects, in one sector. Additionally, there could also be human errors while pressing buttons, i.e incorrect button type, incorrect direction, or unintentional repeat button presses.To address these ambiguities, we developed a systematic approach to manually evaluate button press instances where the corresponding object was not immediately clear. We examined frames both before and after the button press, as well as the participant's gaze history, to identify the most likely object associated with the button press.

B.3 Route & traffic design:

At least one safety critical scenario such as a jaywalking pedestrian was included in each route. We did so to ensure that driver gaze before and during safety critical scenarios was also represented in the dataset. These types of critical scenarios were included:

1. Visible jaywalking pedestrian: A pedestrian visible without occlusions jaywalks into the ego vehicles path.
2. Simultaneous vehicle turning and jaywalking pedestrian: A vehicle turns left or right while entering at an intersection opposite the ego-vehicle. A pedestrian jaywalks behind the turning vehicle.
3. Occluding object jaywalking pedestrian: A pedestrian, visible from afar but occluded as the ego-vehicle nears, jaywalks into the ego vehicles path.
4. Bicycle crossing after turn: Right after the ego-vehicle makes a right turn, a bicyclist crosses the road in front of the ego vehicle
5. Emergency vehicles distracting from pedestrians: Emergency vehicles are parked near a residence. A policeman, partially occluded by a vehicle, jaywalks to the residence.

See the attached video for examples of critical scenarios.

C Modeling Driver SA

Data representation details: The virtual camera used to generate visual sensor data for our model was fixed to be 1.3m above and 1.3m in front of the ego vehicle (measured from the center of the vehicle base). The camera had a 90° field of view and produced 800×600 images.

Model and training details: We used a Feature Pyramid Network [17] segmentation model with a MobileNetV2 [18] backbone (pre-trained on ImageNet). The backbone was chosen for its low number of parameters ($2M$) and runtime efficiency. Our training procedure used the Adam optimizer with a starting learning rate of 10^{-4}. The learning rate was scheduled to drop every 5 epochs by a factor of 5.

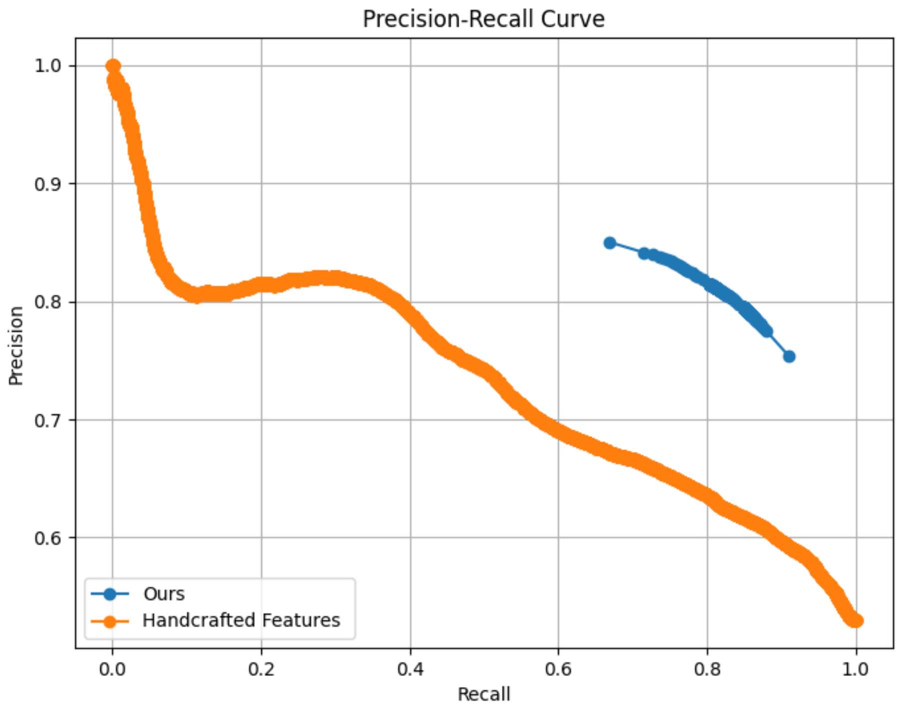


Figure 3: The Precision-Recall curve for our method and the handcrafted-baseline [3]. For our method, we take the mode of predictions over all pixels pertaining to the object, to get the final prediction for the object. To generate the PR curve, our predictions can be thresholded at two levels. First on the raw pixel-level predictions, and second on the ratio of the predicted aware and unaware pixels for a object. Thus, the first threshold level decides what should be the predicted score of a pixel inorder to classify it as aware or unaware. The second level decides how many pixels should be classified as aware inorder to classify this object as aware. To generate this curve we vary the threshold of the raw-pixel level predictions and the second level threshold is fixed at 1. Due to these two levels of thresholds, our method does not have precision = 1 or recall = 1.

Model Ablation	Acc.	Prec.	Recall
Trained from Scratch	76.35%	0.80	0.73
DeeplabV3	69.88%	**0.84**	0.53
Handcrafted Features [3]	65.47%	0.66	0.69
Ours (Full)	**79.21%**	0.83	**0.77**

Table 2: Additional ablations for our model

D Additional Results

A PR curve corresponding to the results in Table 1 in the main paper is shown in Fig. 3. We show two additional baselines in Table 2. We show the effect of pre-training the backbone on ImageNet by comparing it with a network we trained from scratch. The model from scratch was trained with an initial learning rate $10\times$ higher but with the same decaying schedule. We also show results with replacing the Feature Pyramid Network with a DeeplabV3 [19], but it tends to perform about 10% worse. Unet models have been known to perform better for medical image segmentation where target objects are small and the background pixels dominate images [20]. Since our dataset has similar characteristics, this is an expected result.

References

[1] M. R. Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting*, volume 32, pages 97–101. Sage Publications Sage CA: Los Angeles, CA, 1988.

[2] M. R. Endsley, D. J. Garland, et al. Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 1(1):3–21, 2000.

[3] H. Zhu, T. Misu, S. Martin, X. Wu, and K. Akash. Improving driver situation awareness prediction using human visual sensory and memory mechanism. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6210–6216. IEEE, 2021.

[4] D. Sirkin, N. Martelaro, M. Johns, and W. Ju. Toward measurement of situation awareness in autonomous vehicles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 405–415, 2017.

[5] X. Gao, X. Wu, S. Ho, T. Misu, and K. Akash. Effects of augmented-reality-based assisting interfaces on drivers' object-wise situational awareness in highly autonomous vehicles. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 563–572. IEEE, 2022.

[6] F. T. Durso, C. A. Hackworth, T. R. Truitt, J. Crutchfield, D. Nikolic, and C. A. Manning. Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, 6(1):1–20, 1998.

[7] M. R. Endsley. Situation awareness global assessment technique (sagat). In *Proceedings of the IEEE 1988 national aerospace and electronics conference*, pages 789–795. IEEE, 1988.

[8] J. C. de Winter, Y. B. Eisma, C. Cabrall, P. A. Hancock, and N. A. Stanton. Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work*, 21(1):99–111, 2019.

[9] R. M. Taylor. Situational awareness rating technique (sart): The development of a tool for aircrew systems design. In *Situational awareness*, pages 111–128. Routledge, 2017.

[10] M. R. Endsley, S. J. Selcon, T. D. Hardiman, and D. G. Croft. A comparative analysis of sagat and sart for evaluations of situation awareness. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 42, pages 82–86. SAGE Publications Sage CA: Los Angeles, CA, 1998.

[11] A. Kaur, R. Chaujar, and V. Chinnadurai. Effects of neural mechanisms of pretask resting eeg alpha information on situational awareness: a functional connectivity approach. *Human Factors*, 62(7):1150–1170, 2020.

[12] G. Sun, X. Wanyan, X. Wu, and D. Zhuang. The influence of hud information visual coding on pilot's situational awareness. In *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 1, pages 139–143. IEEE, 2017.

[13] W. Liu and W. Xue. An analysis of situation awareness for the car cab display interface assessment based on driving simulation. In *Electronics, Communications and Networks IV*, pages 943–948. CRC Press, 2015.

[14] T. Zhang, J. Yang, N. Liang, B. J. Pitts, K. O. Prakah-Asante, R. Curry, B. S. Duerstock, J. P. Wachs, and D. Yu. Physiological measurements of situation awareness: a systematic review. *Human factors*, page 0018720820969071, 2020.

[15] G. Silvera, A. Biswas, and H. Admoni. Dreyevr: Democratizing virtual reality driving simulation for behavioural & interaction research. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 639–643, 2022.

[16] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[19] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[20] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.