1 Experimental Setup

To assess our proposed model and cleansing method, we utilized the BDD-100K dataset [8] to train the YoloV5 [3] as the object detection model in our cleansing method. The BDD-100K dataset provides labels for 10 object types related to driving tasks, including pedestrian, rider, car, truck, bus, train, motorcycle, bicycle, traffic light, and traffic sign. Additionally, we explored various combinations of object detection datasets and models in the external study. We resized all inputs into 1280×720 which is aligned with the initial size of BDD-A [7] input.

Parameter settings are indicated in Table 1. Our code implementation is based on Pytorch [5] and used the Adam optimizer [4]. We conducted all experiments on an RTX 3090 GPU, while on-board experiments were performed on a Jetson Nano 4G.

Table 1: Parameter Settings

Parameter	**Value**
Token number	16×16
Channel number in CNN	16
Learning rate	0.001
Learning decay	0.1 per 25 epochs
Training epoch	50

2 User Study

To qualitatively determine if our cleansed datasets produce more reasonable human gaze maps than the baseline dataset, we conducted a user study comparing BDD-A [7] with CUEING-B, and DADA-2000 [2] with CUEING-D. BDD-A [7] and CUEING-B contain 30,073 images, while DADA-2000 [2] and CUEING-D contain 22,171 images. To ensure a 95% confidence level with a 10% confidence interval, we randomly sampled 100 images with gaze maps from both the original datasets and their cleansed versions, respectively. In these sampled groups, to obtain quantitative measurements, we asked users to determine how many clusters of gaze each image contains and how many of them are reasonable (the sample user study interface can be found in the supplementary material). We recruited 10 workers from different cultures in Amazon mTurk [1] to participate in the user study, and all of them are car owners and have extensive driving experience. They can receive $1 for each survey they take. We made 5 surveys in total. During the user study, we did not instruct users on what kinds of gazes are reasonable. Users need to rely entirely on their subjective perceptions to determine which image contains a more reasonable gaze.

In Figure 1, the left image has 4 clusters of gaze, and 1 of them is unreasonable, and the right image has 2 clusters of gaze, and all of them are reasonable.

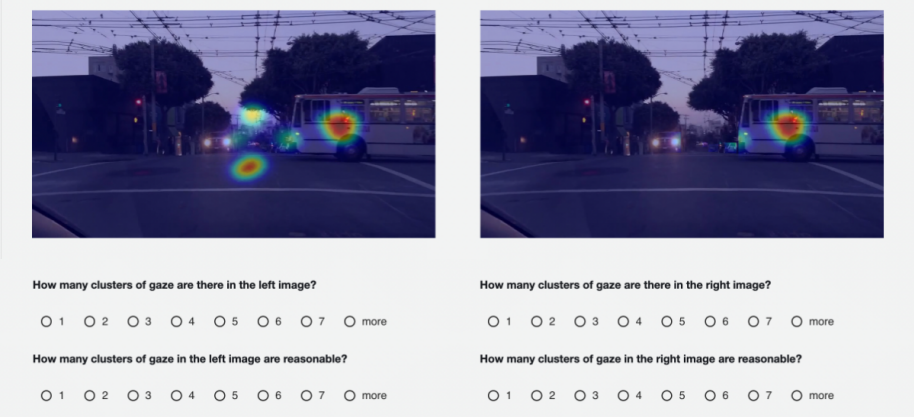

Figure 1: Example of user study interface, the left side is the overlay of image and gaze map from DADA-2000, the right side is the overlay of image and gaze map from CUEING-D.

3 Focus Threshold

Following the approach in [6], if the maximum predicted gaze value within the bounding box exceeds 0.9 (empirically indicating strong focus[1]), the object is considered focused. This threshold is independent of training and used primarily for object-level evaluation.

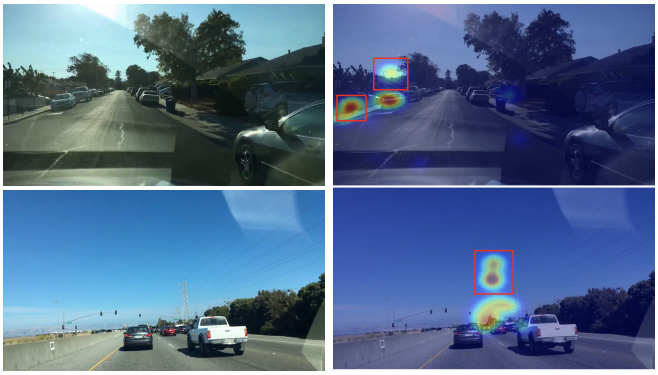

Figure 2: Camera images from BDD-A (first column), Gaze map from BDD-A (second column), red bounding boxes indicate irrelevant items focused by drivers.

[1]During driving, the driver primarily focuses on certain key objects, while peripheral vision may also include other objects. The term "strong focus" refers to the major objects that the driver concentrates on. For example, in Figure 2, the red area in the gaze map indicates the objects that are strongly focused on by the driver.

4 Positional Encoding

To obtain accurate positional information on tokens, we prefer that the location information could be feature-like information in the original input. Therefore the location information can be further processed as a feature by our token convolutional layers in the following stage. Here we employ a specifically designed absolute positional encoding, and assign a two-dimensional coordinate (x, y) to each token. These coordinates will equally spread between -1 and 1 through the rows and columns, which can be denoted as:

$$x = -1 + \frac{2i}{H-1}, \quad y = -1 + \frac{2j}{W-1} \tag{1}$$

where $i \in \{0, \ldots, H-1\}$ and $j \in \{0, \ldots, W-1\}$.

Hence, we can ensure that each token in the original input has a unique coordinate. However, to allow the positional information to be the same dimension as the features of the tokenized input, we use a convolutional layer to map all two-dimensional encoded coordinates from $\mathbf{P} \in \mathbb{R}^{2\times H\times W}$ to $\mathbf{P'} \in \mathbb{R}^{1\times H\times W}$. Then, we perform concatenation between the positional encoding and those tokens in the channel dimension.

5 Loss Function

Equation 2 indicates the loss function, where y is the downsampled ground truth, $\hat{y}$ is the output of the linear layer.

$$L(\hat{y}, y) = -\frac{1}{T}\sum_{i=1}^{T} y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot(1-\log(\hat{y}_i)) \tag{2}$$

References

[1] Crowston, K.: Amazon mechanical turk: A research tool for organizations and information systems scholars. In: Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13-14, 2012. Proceedings. pp. 210–221. Springer (2012)

[2] Fang, J., Yan, D., Qiao, J., Xue, J., Wang, H., Li, S.: Dada-2000: Can driving accident be predicted by driver attentionf analyzed by a benchmark. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 4303–4309. IEEE (2019)

[3] Jocher, G., Stoken, A., Borovec, J., Chaurasia, A., Changyu, L., Laughing, A., Hogan, A., Hajek, J., Diaconu, L., Marc, Y., et al.: ultralytics/yolov5: v5. 0-yolov5-p6 1280 models aws supervise. ly and youtube integrations. Zenodo **11** (2021)

[4] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[5] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)

[6] Rong, Y., Kassautzki, N.R., Fuhl, W., Kasneci, E.: Where and what: Driver attention-based object detection. Proceedings of the ACM on Human-Computer Interaction **6**(ETRA), 1–22 (2022)

[7] Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., Whitney, D.: Predicting driver attention in critical situations. In: Asian conference on computer vision. pp. 658–674. Springer (2018)

[8] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)