MEDIRL: Predicting the Visual Attention of Drivers via Maximum Entropy Deep Inverse Reinforcement Learning

Sonia Baee¹, Erfan Pakdamanian¹, Inki Kim², Lu Feng¹, Vicente Ordonez³, Laura Barnes¹
¹University of Virginia, ²University of Illinois at Urbana Champaign, ³Rice University
sb5ce@virginia.edu, ep2ca@virginia.edu, inkikim@illinois.edu
Lu.feng@virginia.edu, vicenteor@rice.edu, lb3dp@virginia.edu

Abstract

Inspired by human visual attention, we propose a novel inverse reinforcement learning formulation using Maximum Entropy Deep Inverse Reinforcement Learning (MEDIRL) for predicting the visual attention of drivers in accident-prone situations. MEDIRL predicts fixation locations that lead to maximal rewards by learning a task-sensitive reward function from eye fixation patterns recorded from attentive drivers. Additionally, we introduce EyeCar, a new driver attention dataset in accident-prone situations. We conduct comprehensive experiments to evaluate our proposed model on three common benchmarks: (DR(eye)VE, BDD-A, DADA-2000), and our EyeCar dataset. Results indicate that MEDIRL outperforms existing models for predicting attention and achieves state-of-the-art performance. We present extensive ablation studies to provide more insights into different features of our proposed model.¹

1. Introduction

Autonomous vehicles have witnessed significant advances in recent years. These vehicles promise better safety and freedom from the prolonged and monotonous task of driving. However, one of the remaining safety challenges of vision-based models integrated into these vehicles is how to quickly identify important visual cues and understand risks involved in traffic environments at a time of urgency [51]. Humans have an incredible visual attention ability to quickly detect the most relevant stimuli, to direct attention to potential hazards in complex situations [43], and to select only a relevant fraction of perceived information for more in-depth processing [53]. Humans are able to guide their attention by a combination of bottom-up (stimuli driven, e.g., color and intensity) and top-down (task driven, e.g., current goals or intention) mechanisms [13, 27].

¹The code and dataset are provided for reproducibility in https://github.com/soniabaee/MEDIRL-EyeCar.

During task-specific activities, the goal-directed behavior of humans along with their underlying target-based selective attention, enables drivers to ignore objects and unnecessary details in their field of view that are irrelevant to their decisions [7, 8]. For example, at one moment, a driver’s goal might be to initiate an overtaking maneuver, thus a nearby vehicle becomes the target object. Later, the driver may need to stop abruptly to avoid an accident, thereby the brake light of the car in front becomes the target object. Despite recent progress in computer vision models for autonomous systems [28, 63], they are still behind the foveal vision ability of humans [42, 61, 69].

Inverse reinforcement learning (IRL) algorithms are capable to address this problem by learning to imitate the efficient attention allocation produced by an expert, i.e. an attentive driver [41]. It is important that autonomous vehicles leverage human visual attention mechanisms to improve their performance, especially for better safety in critical situations where rare events can be encountered. In this paper, we introduce Maximum Entropy Deep Inverse Reinforcement Learning (MEDIRL) to learn task-specific visual...
attention policies to reliably predict attention in imminent rear-end collisions.

Prior efforts in bottom-up saliency models commonly prioritize pixel location (e.g., free-viewing fixation) [31, 44, 49]. These models do not fully capture driver attention in goal-directed behavior [15, 61, 61, 32]. Moreover, video-based saliency models usually aggregate spatial features guided by saliency maps in each frame [57, 26, 25, 64]. However, most of these fixation prediction models utilized a particular source of information [61, 45, 17], and did not consider to jointly process spatial and temporal information [57, 25]. In this work, we aim to predict eye fixation patterns made prior to critical situations, where these patterns can be either spatial (fixation map) or spatiotemporal (fixation sequences) features.

Inverse reinforcement learning (IRL) is an advanced form of imitation learning [74, 60] that enables a learning agent to acquire skills from expert demonstrations [52]. Our proposed MEDIRL model learns a sequence of eye fixations by considering each fixation as a potential source of reward [65]. We leverage collective visual information that has been deemed relevant for video saliency in prior works [39, 44, 9]. For example, if an autonomous system tries to locate the salient regions of a driving scene before an imminent rear-end collision, the desired visual behavior can be demonstrated by studying the attention of a driver who effectively detects brake lights. In this way, the learning agent can infer a reward function explaining experts’ behavior and optimize its own behavior accordingly. To this end, our proposed model predicts driver attention where a fixation pattern is represented as state-action pairs. Given a video frame input paired with eye fixations, MEDIRL predicts a maximally-rewarding fixation location (action) by perceptually parsing a scene to extract rich visual information (environment), and accumulating a sequence of visual cues through fixations (state) (see Figure 1).

Additionally, we introduce EyeCar, a new driver attention dataset in accident-prone situations. EyeCar is essential for training goal-directed attention models as it is the only dataset capturing attention before accidents in an environment with high traffic density. We exhaustively evaluate our proposed model on three common benchmarks (DR(eye)VE [45], BDD-A [62], DADA-2000 [17]) as well as our own EyeCar dataset. The experimental results show that MEDIRL outperforms state-of-the-art models on driver attention prediction. We also conduct extensive ablation studies to determine which input features are most important for driver attention prediction in critical situations.

Our contributions can be summarized as follows:

- We propose MEDIRL, a novel IRL formulation for predicting driver visual attention in accident-prone situations. MEDIRL uses maximum entropy deep inverse reinforcement learning to predict maximally-rewarding fixation locations.
- We introduce EyeCar, a new driver attention dataset comprised of rear-end collisions videos for the goal-directed attention problem in critical driving situations.
- Extensive experimental evaluation on three driver attention benchmark datasets: DR(eye)VE [45], BDD-A [62], DADA-2000 [17], and EyeCar. Results show that MEDIRL outperforms existing models for attention prediction and achieves state-of-the-art performance. Besides, we present ablation studies showing target (brake light), non-target (context), and driving tasks are important for predicting driver attention.

2. Related Work

Our work is broadly related to prior efforts on models for fixation prediction, using inverse reinforcement learning for visual tasks, and prior datasets for driving tasks.

**Fixation Prediction.** With increased access to large-scale annotated attention datasets and advanced data-driven machine learning techniques, prediction of human saliency has received significant interest in computer vision [59, 56, 31, 73, 11, 39]. A large number of previous studies explored bottom-up saliency models and visual search strategies over static stimuli [16, 34, 22, 18, 4, 67], and video [73, 58, 39, 68, 9]. Generally, the output of these models is an attention map showing the probability of eye fixation distribution. In contrast to this approach, fewer works explored top-down attention models for explaining sequences of eye movements [48, 5, 3]. More recently, some works explored visual attention models in the context of driving [23, 61, 19]. Because task-specific instructions may change gaze distributions [47], some models commonly detect salient regions of images or videos in a free-viewing task. Prior research also studied the pattern of eye movements associated with the task-specific activities [38, 1]. Some of these works rely on the direct ties between eye movement and the demands of a task [65, 50, 48]. These previously proposed attention models are trained mostly on static image-viewing scenarios while human attention typically gets information in a sequential fashion. Further, recent video-saliency works have proposed joint bottom-up and top-down mechanisms for attention prediction using deep learning [45, 62, 17, 29, 44]. However, they did not consider to jointly process spatial and temporal information. We are interested in detecting the salient regions of a scene in a task-specific driving activity in which estimating where the drivers are dynamically looking at, and reliably detecting the task-related objects (target objects).

**Inverse Reinforcement Learning.** Our approach builds on works on modeling human visual attention with their fixation being a sequential decision process of the agent to detect salient regions [37, 70, 35]. The recently proposed work by Yang et al. [65] is the closest to our work as it proposes
Table 1: Compared to prior datasets, EyeCar is the only dataset that captured collisions from a point-of-view (POV) perspective and the host vehicle is involved in the collision. Previous datasets either did not capture attention from a collision point of view or had a less crowded scene. Note that #vehicles refer to the average number of vehicles per frame.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>collision</th>
<th>collision-POV</th>
<th>speed</th>
<th>GPS</th>
<th>#vehicles</th>
<th>#frames</th>
<th>figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DR(eye)VE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>1.0</td>
<td>55k</td>
<td>3</td>
</tr>
<tr>
<td>BDIF-A</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4.4</td>
<td>315k</td>
<td>45</td>
</tr>
<tr>
<td>DADA-2000</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2.1</td>
<td>658k</td>
<td>20</td>
</tr>
<tr>
<td>EyeCar</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4.6</td>
<td>318k</td>
<td>8</td>
</tr>
</tbody>
</table>

3. Method

We propose MEDIRL for predicting drivers’ visual attention in accident prone situations from driving videos paired with their eye fixations. MEDIRL learns a visual attention policy from demonstrated attention behavior. We formulate the problem as the learning of a policy function that models the eye fixations as a sequence of decisions made by an agent. Each fixation pattern is predicted given the present agent state and the current observed world configuration (i.e., a scene context).
sider the road lanes along with the lead vehicle features in our visual representation. The road lanes (\(G_t\)) are critical for the task-related visual attention of drivers since they are an important indicator of the type of maneuver [14]. To amplify the predicted attention for pixels of the target objects, we detect the lead vehicle (\(M_t\)) which is important in rear-end collisions [36]. The lead vehicle is a critical anchor object that can direct the driver attention to the target object, i.e., brake lights. We discretize each frame into an \(n \times m\) grid where each patch matches the smallest (furthest) size of the lead vehicle bounding box (see Figure 2). In addition, we extract pixel locations of the brake lights by first converting each frame to the HSV color space, and then using a position-sensitive ROI max-pooling layer to extract region features for the lead vehicle box (\(U_l\)). The boxes and their respective features are treated as a set of objects.

**Relative Distance.** Drivers pay more attention to the objects which are relatively closer as opposed to those at a distance, since the chance of collision is significantly higher for the former case. Thus, relative distance between objects and the ego-vehicle is crucial for making optimal driving decisions [44]. To amplify nearby regions of a driving scene, we use dense depth map (\(D_t\)) and combine it with the general visual features (\(Y_t\)) by using the following formula:

\[
Z_t = Y_t \oplus D_t = Y_t \circ \lambda \ast D_t + Y_t, \lambda = 1.2
\]

where \(\lambda\) is an amplification factor.

**Driving Tasks.** To discover which features of an observed environment are the most driving task related, we need to determine the types (\(Q_t\)) of driving task. We observed three driving tasks ending to rear-end collisions across all videos: lane-keeping, merging-in, and braking. We use function \(f_{task}\) to define these driving tasks by two criteria: 1) ego-vehicle makes lane changing decision \(c\) and 2) the existence of a traffic signal \(I_{signal}\) in a given driving video.

\[
\text{driving task} = \begin{cases} 
\text{lone-keeping,} & \text{if } c = 0, \ I_{signal} = 0 \\
\text{merging-in,} & \text{if } c = 1, \ I_{signal} = 0 \text{ or } 1 \\
\text{braking,} & \text{if } c = 0, \ I_{signal} = 1 
\end{cases}
\]

**Vehicle State.** We optionally concatenate the speed of the ego-vehicle \(v_t\), which can influence the fixation selection [66, 45, 44], with the extracted visual representation, relative distance, and driving tasks.

### 3.2. MEDIRL

Attentive drivers predominantly attend to the task-related regions of the scene to filter out irrelevant information and ultimately make the optimal decisions. MEDIRL attempts to imitate this behavior by using the collective non-target and target features extracted through parsing the driving scene in the state representation. Subsequently, it integrates changes in the state representation with alterations in eye fixation point, to predict fixation. Therefore, the **state** of an agent is determined by a sequence of visual information that accumulates through fixations towards the target object (i.e., a brake light) which we call it a foveated frame. Figure 1 shows an example of a foveated frame. The **action** of an agent, the next fixation location, depends on the state at that time. The **goal** of an agent is to maximize internal **reward** by encapsulating the intended behavior of attentive drivers (experts) through changes in fixation locations. MEDIRL employs IRL to recover this reward function (\(R\)) from the set of demonstrations.

**State Representation:** MEDIRL considers the following components in the state representations: simulating the human visual system, collecting a context of spatial cues, and modeling state dynamics. See Algorithm 1 for describing the overview of the state representation.

**Human visual system (fovea):** Human visual system accumulates information by attending to a specific location within the field of view. Consequently, humans selectively fixate on new locations to make optimal decisions. It means high-resolution visual information is available only at a central fixated location and the visual input outside of the attend location becomes progressively more blurred with distance away from the currently fixated location [69]. We simulate human fovea by capturing high-resolution information about the current fixation location and a surrounding patch with a size 12 \(\times\) 17 (about 1° visual angle), as well as low-resolution information outside of the simulated fovea [69]. To effectively formulate this system, MEDIRL uses a local patch from the original frames of the video as the high-resolution foveal input and a blurred version of the frame to approximate low-resolution input \(L\) from peripheral vision [71]. We obtain the blurred frames by applying a Gaussian smoothing with standard deviation \(\sigma = 2 \times d\), which \(d\) is equal to Euclidean distance between the current fixation point \(p_{x,t}\), where \(k = 0, ..., K\), and the size of the frame. Note that the number of fixations \(K\) varies from frame to frame.

**Spatial cues:** A driving task and the driving-relevant (anchor) objects of the scene can potentially direct drivers’ attention to the primary target object. For example, drivers consider the distance to the lead vehicle when they brake. To approximate this guided selection of fixations, MEDIRL includes visual information in the state representation. This state representation collects the non-target and target features can create a context of spatial and temporal cues that might affect the selection of drivers’ fixations.

**Dynamics of state:** To model the altering of the state representation followed by each fixation, we propose a dynamic state model. To begin with, the state is a low-resolution frame corresponding to peripheral visual input. After each fixation made by a driver, we update the state by replacing the portion of the low-resolution features with
Algorithm 1 MEDIRL State Representation

```
1: function VISUAL_ENCODER(a video frame I)
2:   X := HRnet(I) \hspace{1cm} \triangleright\ global feature
3:   O := mask-rcnn(I) \hspace{1cm} \triangleright\ list of detected object
4:   Y := ROAveragexO(X) \hspace{1cm} \triangleright\ extract region features
5:   G, c := VGG-net(I) \hspace{1cm} \triangleright\ detect road lanes and lane changes
6:   M, I_{signal} := mask-rcnn(Y) \hspace{1cm} \triangleright\ detect lead-vehicle and traffic signal
7:   U := ROI-max(HSV-color)(I, M) \hspace{1cm} \triangleright\ detect brake lights
8:   D := MonoDepth2(I) \hspace{1cm} \triangleright\ compute relative distance
9:   Z := Y \oplus D \hspace{1cm} \triangleright\ amplify close objects
10:  Q := f_{\text{smooth}}(c, I_{signal}) \hspace{1cm} \triangleright\ compute driving task
11:  v := concatenate(G, M, U, Z) \hspace{1cm} \triangleright\ a context of spatial cues
12:  return visual-cues, v, \tau
13: function BLUR(frame I, fixation k)
14:   d := Euclidean(k, size(I))
15:   \Gamma := GaussianBlur(1, \sigma) \hspace{1cm} \triangleright\ apply a Gaussian smoothing
16:   return I
17: function STATE_DYNAMICS(frame I_k, fixations K)
18:   for k \in K do
19:     # collect context of spatial cues based on a simulated fovea movements
20:     H_k := VisualEncoder(I_k)
21:     L_{k,1} := VisualEncoder(blur(I_k))
22:     # update the state that occurs following each fixation
23:     O_{k,1} := L_{k,1} \hspace{1cm} \triangleright\ initialize frame corresponding to peripheral vision
24:     # \mathbf{E}_{k,1}^s is the circular mask generated from the kth fixation
25:     # return all spatial six cues of the selected patch in the frame is a new fixation
26:     O_{k+1,1} = E_{k,1} \odot H_k + (1 - E_{k,1}) \odot O_{k,t}
27:     where \odot is an element-wise product. O_{k,t} is a context of spatial cues after k fixations. E_{k,t} is the circular mask generated from the kth fixation (i.e., it is a binary map with 1 at current fixation location and 0 elsewhere in a discretize frame). To jointly aggregate all the temporal information, we update the next frame by considering all context of spatial cues in the previous frame as follows:
28:     O_{k,t+1} = E_{k,t+1} \odot H_{k+1} + (1 - E_{k,t+1}) \odot O_{K,t}
29:     where O_{K,t} is visual information after all fixations K of time step t(previous frame).
30:   Drivers have various visual behaviors while performing a driving tasks and many factors (e.g. speed) may affect the chosen fixation strategy [66, 45, 44]. To efficiently predict fixations for all drivers, we augment the state by aggregating it with a high-dimensional latent space that encodes the driving task Q_i. We then add another fully-connected layer to encode the current speed of the ego-vehicle v_t and concatenate the state with the speed vector. With the visual information and ego-vehicle state at each time step, we fuse all into a single state. The state of the agent is then complete in the sense that it contains all bottom-up, top-down, and historical information (more detail of these components can be found in the supplementary material).
31: Action Space: Herein we aim to predict the next eye fixation location of a driver. Therefore, the policy selects one out of n \times m patches in a given discretize frame. The center
32: of the selected patch in the frame is a new fixation. Finally, the changes (\Delta_x, \Delta_y) of the current fixation and the selected fixation define the action space A_t: \{left, right, up, down, focus-inward, focus-outward, stay\}, as shown in Figure 1 which has three degrees of freedom (vertical, horizontal, diagonal).
33: Reward and Policy: To learn the reward function and policies of driver visual attention in rear-end collisions, we use a maximum entropy deep inverse reinforcement learning [60]. MEDIRL assumes the reward is a function of the state and the action, and this reward function can be jointly learned using the imitation policy.
34: The main goal of IRL is to recover the unknown reward function R from the set of demonstrations \Xi = \{\xi_1, \xi_2, \ldots, \xi_q\}, where \xi_q = \{(s_1, a_1), \ldots, (s_T, a_T)\}. We use maximum entropy deep IRL, which models trajectories as being distributed proportional to their exponentiated return: p(\xi) = (\gamma/z) \exp(R(\xi)),
35: where Z is the partition function, Z = \int \exp(R(\xi))d\xi. To approximate the reward function, we assume it can be represented as R = \omega^T \phi, where \omega is a weight vector \phi is a feature vector. Such representation is constrained to be linear with respect to the input features \phi. In order to learn a reward function with fewer constraints, we use deep learning techniques to determine \Phi(\phi, \theta), a potentially higher dimensional feature space, and approximate the reward function as R = \omega^T \Phi(\phi, \theta)(s, a). Note that the weight vectors of \omega and the parameter vector \theta are both associated with the network which is fine-tuned by jointly training the different category of driving tasks.
36: Loss Function: To learn the attention policies, MEDIRL maximizes the joint posterior distribution of fixation selection demonstrations \Xi, under a given reward structure and of the model parameter, \theta. For a single frame and a given fixation sequence \xi, with a length of |\tau|, the likelihood is:
37: L_\theta = \frac{1}{|\tau|} \sum_{\xi^i \in \Xi} \log P(\xi^i, \theta),
38: where P(\xi^i, \theta) is the probability of the trajectory \xi^i in demonstration \Xi.
39: The algorithm tries to select a reward function that induces an attention policy with a maximum entropy distribution over all state-action trajectories and minimum empirical Kullback-Leibler divergence (KLD) from drivers state-action pairs. In each iteration (q) of maximum entropy deep IRL algorithm, we first evaluate the reward value based on the state features and the current reward network parameters (\theta_q). Then, we determine the current policy (\pi_q) based on the current approximation of reward (R_q), and transition matrix T (i.e., the outcome state-space of a taken action). We benefit from the maximum entropy paradigm, which enables the model to handle sub-optimal and stochastic visual
```
The simulator consisting of a Logitech G29 steering wheel, accelerator, brake pedal, and eye-tracker (see supplementary materials for more details).

We recruited 20 participants (5 female and 15 male, ages 22-39) with at least three years of driving experience (Mean=9.7, SD=5.8). Participants watched all 21 selected dash-cam videos (each lasted approximately 30sec) to identify hazardous cues in rear-end collisions. The EyeCar dataset contains 3.5 hours of gaze behavior (aggregated and raw) captured from more than 315,000 rear-end collisions video frames. Each frame comprises 4.6 vehicles on average which makes EyeCar driving scenes more complex than other visual attention datasets (see Table 1). The extracted speed from each frame shows that 38% of vehicles were driving high (\(v \geq 35\)), 39% normal (\(35 < v \leq 65\)), and 23% low (\(v \leq 35\)). EyeCar also provides a rich set of annotations(e.g., scene tagging, object bounding, lane marking, etc.; more details in supplementary materials).

### 5. Experiments

**Training details.** Driver attention is often strongly biased towards the vanishing point of the road and does not regularly change in a normal driving situation [62, 44]. However, attentive drivers regularly shift their attention from the center of the road to capture important cues in accident-prone situations. MEDIRL aims to predict driver attention in critical situations. Thus, to learn driving task-specific fixations and to avoid a strong center bias in our model two criteria were imposed when sampling training frames: 1) train on important frames, 2) exclude driving-irrelevant objects fixation sequence. Since a driver has to attend (fixate) to important visual cues which usually appear in critical situations, the important frames are defined as frames wherein the attention map greatly deviates from the average attention map. We use KLD to measure the difference between
After applying the exclusion standard, we were left with BDD-A, and rear-end collision events from DADA-200.

Datasets. We evaluate our model on three driver attention benchmark datasets: DR(eye)VE [45], BDD-A [62], DADA-2000 [17] and EyeCar. To predict driver attention related to rear-end collisions, we extract the full stopping gaze patterns of all independent observers [12]. We then sample continuous sequences of six frames as the training frames where their KLD is at least 0.89. We also exclude fixation sequences with more than 40% focus on the irrelevant objects (e.g., trees, advertisement).

We resize each video frame input to 144 × 256. Then we normalize each frame by subtracting the global mean from the raw pixels and dividing by the global standard deviation. To encode visual information (see Sec. 3), we use several backbones: HRNetV2 [55]–pre-trained on Mapillary Vistas street-view scene [40], MaskTrack-RCNN [64]–pre-trained on youtube-VIS, Monodepth2 [21]–pre-trained on KITTI 2015 [20], and VPGNet [33]–pre-trained on VPGNet dataset.

MEDIRL consists of four hidden convolutional layers with 52, 34, 20, and 20 ReLu units, respectively; followed by seven softmax units to output a final probability map. We use batch normalization after ReLu activation and set the reward discount factor to 0.98. We also set the initial learning rate to 1.5 × 10⁻⁴, and during the first 10 epochs, we linearly increase the learning rate to 5 × 10⁻⁴. After epoch 11, we apply a learning rate decay strategy that multiplies the learning rate by 0.25 every three epochs. For training, we use Adam optimizer [30] (β₁ = .9, β₂ = .99) and weight decay = 0. Overall, MEDIRL is trained on 36 hours on a single NVIDIA Tesla V100 GPU and it takes about 0.08 seconds to process each frame.

Evaluation Metrics. To evaluate attention prediction, we use several evaluation metrics, including the average attention map of the entire video. The average attention map of each frame is calculated by aggregating and smoothing the gaze patterns of all independent observers [12]. We then sample continuous sequences of six frames as the training frames where their KLD is at least 0.89. We also exclude fixation sequences with more than 40% focus on the irrelevant objects (e.g., trees, advertisement).

Datasets. We evaluate our model on three driver attention benchmark datasets: DR(eye)VE [45], BDD-A [62], DADA-2000 [17] and EyeCar. To predict driver attention related to rear-end collisions, we extract the full stopping events (resembling near-collisions) from DR(eye)VE and BDD-A, and rear-end collision events from DADA-200.

Table 3: Performance comparison of driver attention prediction on EyeCar. The models trained on Dr(eye)VE [45], BDD-A [62], and DADA-2000 [17] train sets and tested on EyeCar.

![Image](https://via.placeholder.com/150)

Figure 3: Predicted driver attention in a braking task for each compared model and MEDIRL. They all trained on BDD-A. MEDIRL can learn to detect most task-related salient stimuli (e.g., traffic light, brake light). Redder color indicates the expectation of higher reward for fixation location. More examples in supplementary materials.
use location-based and distribution-based saliency metrics: KLD, shuffled Area under the ROC curve (s-AUC), and Correlation Coefficient (CC) [6]. We report s-AUC since it penalizes models with more central prediction [5, 6, 19].

### 6. Results

Table 2 provides the quantitative evaluation results of MEDIRL and five baseline attention prediction models including Multi-branch [45], HWS [62], SAM-ResNet [11], SAM-VGG [11], TASED-NET [39]. For fair comparisons, we directly report available results released by the authors or reproduce experimental results via publicly available source codes. In this evaluation, we trained models on BDD-A and tested on each benchmark. The results highlight that MEDIRL surpasses almost all models under all evaluation metric. Most significantly, our approach can effectively predict driver attention while performing various driving tasks. Although we are unable to calculate s-AUC for Dr(eye)VE as the original fixation were not reported, the results in Table 2 also indicates that the MEDIRL’s superiority is not limited to a dataset.

Further, we evaluate MEDIRL along with other attention models on EyeCar dataset, reported in Table 3. In this experiment, we trained models on each benchmark (i.e., BDD-A, DR(eye)VE, DADA) and tested on EyeCar. MEDIRL performs favorably against other counterparts. However, there is a big performance gap between Table 2 and 3, which may indicate EyeCar has different distributions. To investigate this matter, we trained models on EyeCar and tested on each benchmark. We obtained the following results: (CC : 0.89, KLD : 0.80), (CC : 0.94, s-AUC : 0.91, KLD : 0.85), (CC : 0.85, s-AUC : 0.77, KLD : 0.99) on DR(eye)VE, BDD-A, and DADA-2000, respectively, that are average values for all types of driving tasks. These results show the effectiveness of EyeCar on representing salient regions in critical situations and also show that EyeCar attention distribution prior to accident-prone situations is more informative than benchmarks.

Figure 3 shows qualitative comparison of MEDIRL against other models. MEDIRL can reliably capture the important visual cues in a braking task in the case of a complex frame. In contrast, nearly all other models partially capture the spatial cues and predict attention mainly towards the center of the frame, thereby ignoring the target and non-target objects (i.e., spatial cues). Please refer to the supplementary material for more examples.

### 6.1. Ablations Studies

To investigate how different features in our model affect its performance, we compare several ablated versions of our model against two testing sets (i.e., EyeCar and BDD-A), using $F_\beta$ ($\beta^2 = 1$ [44]), CC, and KLD. All ablated versions of our model are trained on BDD-A.

The results show that crucial features in the model include the context of spatial cues related to target and non-target (L3), driving-specific objects (Line 8, 10), followed by driving task (L9) features. MEDIRL without target (L2) and non-target (L5) shows a significant performance drop. From the results in Table 4, we can observe that compared with the ablated versions, our full model achieves better performance, which demonstrates the necessity of each feature in our proposed model.

### 7. Conclusion

We proposed MEDIRL, a novel inverse reinforcement learning formulation for predicting driver attention in accident-prone situations. MEDIRL effectively learns to model the fixation selection as a sequence of states and actions. MEDIRL predicts a maximally-rewarding fixation location by perceptually parsing a scene and accumulating a sequence of visual cues through fixations. To facilitate our study, we provide a new driver attention dataset comprised of rear-end collision videos with richly annotated eye information. We investigate the effectiveness of attention prediction model by experimental evaluation on three benchmarks and EyeCar. Results show that MEDIRL outperforms existing models for attention prediction and achieves state-of-the-art performance.

**Acknowledgements** This work was supported in part by a gift from Leidos, National Science Foundation (NSF) grant CCF-1942836, and IIS-2045773.
References


13186


