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

Sonia Baee[1], Erfan Pakdamanian[1], Inki Kim[2], Lu Feng[1], Vicente Ordonez[3], Laura Barnes[1]
[1]University of Virginia, [2]University of Illinois at Urbana Champaign, [3]Rice 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].

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

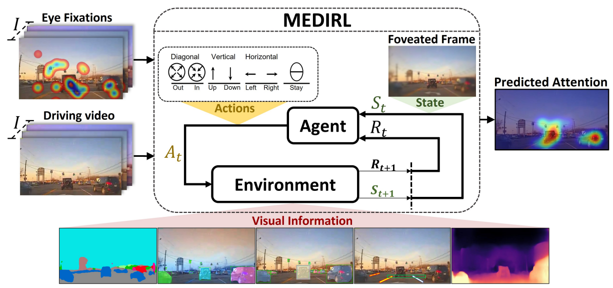

Figure 1: Given a driving video and corresponding eye fixations as inputs, MEDIRL learns to model the fixation selection as a sequence of states and actions (S_t, A_t). MEDIRL then predicts a maximally-rewarding fixation location by perceptually parsing a scene to extract rich visual information (environment) and accumulating a sequence of visual cues through fixations (state).

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, 38, 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

Dataset	collision	collision-POV	speed	GPS	#vehicles	#frames	#gaze
DR(eye)Ve	✗	✗	✓	✓	1.0	555k	8
BDD-A	✗	✗	✓	✗	4.4	318k	45
DADA-2000	✓	✗	✗	✗	2.1	658k	20
EyeCar	✓	✓	✓	✓	4.6	315k	20

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.

a model of visual attention in a visual search task of *static images*. We go further by addressing video saliency predictions in a dynamic and complex driving environment. Our model also does not require to predefine a set of targets but instead parses each driving video frame to extract rich scene context and candidate target objects. Next, it integrates visual cues with driver's eye fixations. It then recovers the intrinsic task-specific reward functions [72] induced by visual attention allocation policies recorded from drivers in a driving environment. To do that, we propose to use maximum entropy deep IRL [74] which can handle raw image inputs and enables the model to handle the often sub-optimal and seemingly stochastic behaviors of drivers [60].

Driving Attention Datasets. Several driving behavior datasets have been proposed [10, 63, 46]. However, only a few large-scale, publicly available, real-world video datasets with annotated visual attention exist in a driving context. DR(eye)VE [45] and BDD-A [62] are the most well-known large-scale annotated datasets in naturalistic and in-lab driving settings, respectively. Importantly, the recently-released annotated driving attention dataset with in-lab settings, DADA-2000 [17], is the only available dataset capturing scenes of collisions. This is because it is nearly impossible to collect enough driver attention data for collision or near-collision events. EyeCar further contributes to this area by having a more diverse array of driving events, beyond looking forward and lane-keeping. Unlike DADA-2000, EyeCar captures collisions from a collision point-of-view (POV) perspective (egocentric) where the ego-vehicle is involved in the accident. Table 1 compares EyeCar with similar datasets (more details in Sec. 4).

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).

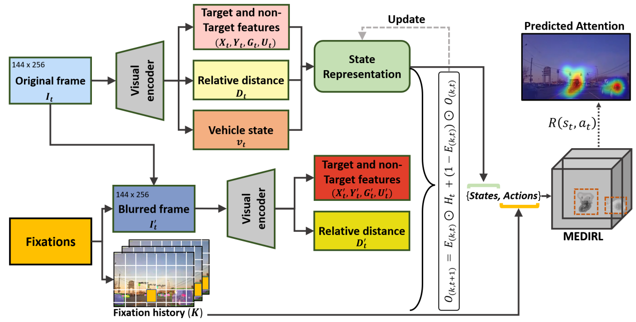

Figure 2: Overview of our state-representation. To simulate human fovea, the agent receives high-resolution information surrounding the attended location, and low-resolution information outside of this simulated fovea. At each fixation point, a new state is generated by applying Eq. 3.2.

3.1. Overview and Preliminaries

In this section, we introduce our notation and describe the features used in our proposed model.

Visual Information. During attention allocation in a dynamic and complex scene, relevant anchor objects–those with a spatial relationship to the target object–can guide attention to a faster reaction time, less scene coverage, and less time between fixating on the anchor and the target object [54, 24, 2]. Therefore, we need to encode each frame of a given video to extract target and non-target features which an agent needs in order to effectively select the next fixation locations. Next, we describe in detail how this encoding is done (see Figure. 2). An overview of the visual encoder function is also outlined in Algorithm 1.

Given a family of driving video frame input, $I = \{I_t\}_{t=1}^{T}$, where T is the number of frames. We extract visual information in a discriminative way while keeping the relevant spatial information. Each frame has several fixation locations that are processed sequentially. At each step, we extracts visual features from the current input frame. To well represent a given video frame input to an agent, we consider both pixel- and instance-level representation (see Figure 1). The pixel-level representation determines the overall scene category by putting emphasis on understanding its global properties. The instance-level representation identifies the individual constituent parts of a whole scene as well as their interrelations on a more local instance-level.

For pixel-level representations, we extract features X_t from a given video frame (e.g., cars, trees). The feature extractor output is a tensor $X_t \in \mathcal{R}^{h \times w \times d}$, where h, w, and d are the height, width, and channel, respectively. At the instance-level, we represent the bounding box or instance-mask to reason explicitly over instances (e.g., lead-vehicle) rather than reasoning over all objects representation. We utilize a position-sensitive ROI average pooling layer [64] to extract region features Y_t for each box.

To extract features relevant to a driving task, we also consider 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_t). 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:

$$Z_t = Y_t \oplus D_t = Y_t \odot \lambda * D_t + Y_t, \lambda = 1.2$$

where λ 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.

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

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 12×17 (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 $p_{k,t}$, where $k = 0, ..., \mathcal{K}$, 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

```
function VISUAL ENCODER(a video frame I)
    X := HRnet(I)                                   ▷ global feature
    O := mask-rcnn(I)                               ▷ list of detected object
    Y := ROI-average(O, X)                          ▷ extract region features
    G, c := VPG-net(I)                              ▷ detect road lanes and lane changes
    M, I_signal := mask-rcnn(Y)                     ▷ detect lead-vehicle and traffic signal
    U := ROI-max(HSV-color(I),M)                    ▷ detect brake lights
    D := MonoDepth2(I)                              ▷ compute relative distance
    Z := Y ⊕ D                                      ▷ amplify close objects
    Q := f_task(c, I_signal)                        ▷ compute driving task
    visual-cues = concatenate(G, M, U, Z)           ▷ a context of spatial cues
    v := ego-vehicle speed                          ▷ vehicle state
    return visual-cues, v, Q                        ▷ return all extracted features
function BLUR(frame I, fixation k)
    d = Euclidean(k,size(I))
    I' = GaussianBlur(I, σ), σ = 2 × d              ▷ apply a Gaussian smoothing
    return I'                                       ▷ return the low-resolution frame
procedure STATE DYNAMICS(frame I_t, fixations K)
    for k ∈ K do do
        # collect context of spatial cues based on a simulated fovea movements
        H_t := VisualEncoder(I_t)
        L_{k,t} := VisualEncoder(blur(I_t,k))
        # update the state that occurs following each fixation
        O_{0,1} = L_{0,1}                           ▷ initialize frame corresponding to peripheral vision
        # E_{k,t} is the circular mask generated from the k-th fixation
        O_{k+1,t} = E_{k,t} ⊙ H_t + (1 − E_{k,t}) ⊙ O_{k,t}
```

the corresponding high-resolution portion obtained at each new fixation location (see Figure. 2). At a given time step t, feature maps H for the original frame (high-resolution) and feature maps L for the blurred frame (low-resolution) are combined as follows:

$$O_{0,1} = L_{0,1}, O_{k+1,t} = E_{k,t} \odot H_t + (1 - E_{k,t}) \odot O_{k,t},$$

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:

$$O_{k,t+1} = E_{k,t+1} \odot H_{t+1} + (1 - E_{k,t+1}) \odot O_{\mathcal{K},t},$$

where $O_{\mathcal{K},t}$ is visual information after all fixations $\mathcal{K}$ of time step t(previous frame).

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_t. 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).

Action Space: Herein we aim to predict the next eye fixation location of a driver. Therefore, the policy selects one out of $n * m$ patches in a given discretize frame. The center of the selected patch in the frame is a new fixation. Finally, the changes (Δ_x, Δ_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).

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.

The main goal of IRL is to recover the unknown reward function R from the set of demonstrations $\Xi = \{\xi_1, \xi_2, ..., \xi_q\}$, where $\xi_q = \{(s_1, a_1), ..., (s_\tau, a_\tau)\}$. We use maximum entropy deep IRL, which models trajectories as being distributed proportional to their exponentiated return:

$$p(\xi) = (1/z) exp(R(\xi)),$$

where Z is the partition function, $Z = \int_\xi exp(R(\xi)) d\xi$. To approximate the reward function, we assume it can be represented as $R = \omega^T \phi$, where ω is a weight vector and ϕ is a feature vector. Such representation is constrained to be linear with respect to the input features ϕ. 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 ω and the parameter vector θ are both associated with the network which is fine-tuned by jointly training the different category of driving tasks.

Loss Function: To learn the attention policies, MEDIRL maximizes the joint posterior distribution of fixation selection demonstrations Ξ, under a given reward structure and of the model parameter, θ. For a single frame and a given fixation sequence ξ with a length of $|\tau|$, the likelihood is:

$$\mathcal{L}_\theta = (1/\Xi) \sum_{\xi^i \in \Xi} logP(\xi^i, \theta),$$

where $P(\xi^i, \theta)$ is the probability of the trajectory ξ^i in demonstration Ξ.

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 (θ_q). Then, we determine the current policy (π_q) based on the current approximation of reward (R_q), and transition matrix $\mathcal{T}$ (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

Data	Method \ Task	Merging-in CC↑	Merging-in s-AUC↑	Merging-in KLD↓	Lane-keeping CC↑	Lane-keeping s-AUC↑	Lane-keeping KLD↓	Braking CC↑	Braking s-AUC↑	Braking KLD↓
DR(eye)VE [45]	Multi-branch [45]	0.48	-	2.80	0.55	-	1.87	0.71	-	2.20
	HWS [62]	0.51	-	2.12	0.75	-	1.72	0.74	-	1.99
	SAM-ResNet [11]	**0.78**	-	2.01	0.80	-	1.80	0.79	-	1.89
	SAM-VGG [11]	**0.78**	-	2.05	0.82	-	1.84	0.80	-	1.81
	TASED-NET [39]	0.68	-	1.89	0.73	-	1.71	0.70	-	1.89
	MEDIRL (ours)	**0.78**	-	**0.88**	**0.89**	-	**0.75**	**0.85**	-	**0.82**
BDD-A [62]	Multi-branch [45]	0.58	0.51	2.08	0.75	0.72	2.00	0.69	0.77	2.04
	HWS [62]	0.53	0.59	1.95	0.67	0.89	1.52	0.69	0.81	1.59
	SAM-ResNet [11]	0.74	0.61	2.00	0.89	0.79	1.83	0.85	0.88	1.89
	SAM-VGG [11]	0.76	0.62	1.79	0.89	0.82	1.64	0.86	0.87	1.85
	TASED-NET [39]	0.73	0.68	1.83	0.81	0.66	1.17	0.87	0.88	1.12
	MEDIRL (ours)	**0.82**	**0.79**	**0.91**	**0.94**	**0.91**	**0.85**	**0.93**	**0.92**	**0.89**
DADA-2000 [17]	Multi-branch [45]	0.44	0.53	3.65	0.69	0.54	2.85	0.67	0.64	2.91
	HWS [62]	0.49	0.59	3.02	0.72	0.53	2.65	0.69	0.77	2.80
	SAM-ResNet [11]	0.65	0.61	2.39	0.78	0.64	2.32	0.75	0.81	2.34
	SAM-VGG [11]	0.68	0.60	2.41	0.76	0.62	2.24	0.75	0.80	2.35
	TASED-NET [39]	0.69	0.66	1.98	0.78	0.69	1.87	0.80	0.81	1.45
	MEDIRL (ours)	**0.70**	**0.68**	**1.31**	**0.89**	**0.71**	**0.92**	**0.81**	**0.88**	**0.99**

Table 2: Performance comparison of driver attention prediction on benchmarks. Models trained on the BDD-A [62] train set and tested on Dr(eye)VE [45], BDD-A [62], and DADA-2000 [17] test sets.

behavior of drivers, by operating on the distribution over possible trajectories [74, 60].

4. The EyeCar Dataset

Attentional lapses in normal situations (e.g., lane-following, empty road) do not cost the same as accident-prone situations (e.g., rear-end collision) where the cost of making an error is high. Nevertheless, collecting enough eye movements from drivers in accident-prone situations is nearly impossible because they are rather uncommon. In addition, driver attention data collected in-car has two main drawbacks [62, 61]: 1) missing covert attention: eye-trackers can only record a single focus of drivers while a driver may be attending to multiple important objects, and 2) false positive gaze: drivers can be distracted to potential disturbances (e.g., side road advertisement) that are not relevant to the driving. Prior works [62, 61] addressed these issues with in-lab data collection, collecting drivers' eye movements while performing simulated driving tasks.

Although in-lab driver attention collection is inevitably different from in-car driver attention, BDD-A in-lab experimental protocol showed that in-lab visual attention data reliably reveal where a driver should look at and identify the potential risks. Therefore, we follow their established and standardized experimental design protocol for collecting in-lab driver attention and create the EyeCar dataset exclusively for rear-end collisions. In order to incentivize users to pay attention and play the fall-back ready role in autonomous vehicles, we further modified the experimental design by sitting them in a low-fidelity driving simulator. 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 ($65 \leq v$), 39% normal ($35 \leq v \leq 65$), and 23% low ($35 \geq v$). 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

Data	Task / Method	Merging-in CC↑	Merging-in s-AUC↑	Merging-in KLD↓	Lane-keeping CC↑	Lane-keeping s-AUC↑	Lane-keeping KLD↓	Braking CC↑	Braking s-AUC↑	Braking KLD↓
DR(eye)VE [45]	Multi-branch [45]	0.36	0.37	6.46	0.51	0.49	4.80	0.69	0.49	3.38
	HWS [62]	0.38	0.34	4.38	0.71	0.51	4.44	0.72	0.61	3.30
	SAM-ResNet [11]	0.49	0.48	4.29	0.73	0.55	3.90	0.74	0.66	3.27
	SAM-VGG [11]	0.50	0.47	4.31	0.74	0.53	3.95	0.75	0.64	3.29
	TASED-NET [39]	0.48	0.46	3.95	0.74	0.55	3.81	0.76	0.65	3.23
	MEDIRL (ours)	**0.51**	**0.51**	**2.32**	**0.76**	**0.57**	**3.11**	**0.79**	**0.69**	**3.07**
BDD-A [62]	Multi-branch [45]	0.46	0.48	4.42	0.51	0.61	3.57	0.61	0.64	3.08
	HWS [62]	0.41	0.47	4.36	0.69	0.81	3.55	0.67	0.68	2.86
	SAM-ResNet [11]	0.55	0.48	3.85	0.85	0.72	3.29	0.79	0.74	2.46
	SAM-VGG [11]	0.53	**0.49**	3.92	0.84	0.70	3.22	0.77	0.70	2.49
	TASED-NET [39]	0.55	**0.49**	3.78	0.84	0.71	3.12	0.77	0.76	2.47
	MEDIRL (ours)	**0.58**	**0.49**	**2.81**	**0.86**	**0.73**	**2.43**	**0.79**	**0.81**	**2.30**
DADA-2000 [17]	Multi-branch [45]	0.21	0.38	6.46	0.45	0.44	4.67	0.54	0.59	3.12
	HWS [62]	0.31	0.35	6.12	0.51	0.47	4.54	0.67	0.71	3.10
	SAM-ResNet [11]	0.33	0.38	5.28	0.65	0.56	4.42	0.77	0.71	3.07
	SAM-VGG [11]	0.30	0.39	5.35	0.69	0.57	4.31	0.74	0.69	3.10
	TASED-NET [39]	0.32	0.38	4.76	0.68	0.57	3.99	0.73	0.74	3.01
	MEDIRL (ours)	**0.41**	**0.45**	**3.79**	**0.73**	**0.60**	**2.51**	**0.75**	**0.79**	**2.51**

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.

the attention over each video frame and 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. After applying the exclusion standard, we were left with 400, 1350, and 534 events in DR(eye)VE, BDD-A, and DADA-200, respectively. Finally, within each type of driving task, we randomly split each of them into three sets of: 70% training, 10% validation, and 20% test.

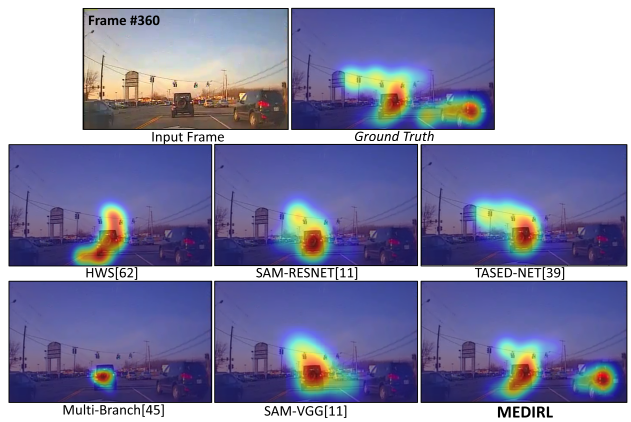

Figure 3: Predicted driver attention in a braking task for each compared model and MEDIRL. They all trained on BDD-A. MEDIRL can learn to detect most task-related salient stimuli (e.g., traffic light, brake light). Redder color indicates the expectation of higher reward for fixation location. More examples in supplementary materials.

5.1. Implementation Details

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] ($\beta_1 = .9, \beta_2 = .99$) and weight decay = 0. Overall, MEDIRL is trained on 36 epochs with a batch size of 20 sequences, and each sequence had six frames. The training time of MEDIRL is approximately 1.5 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

	Dataset / Ablated versions of MEDIRL	EyeCar			BDD-A [62]		
		CC↑	KLD↓	F_β ↑	CC↑	KLD↓	F_β ↑
1	global image + IRL	0.18	4.21	0.10	0.22	4.38	0.12
2	non target + IRL	0.19	4.15	0.12	0.20	4.29	0.12
3	target+non target + IRL	0.29	3.51	0.18	0.36	3.85	0.25
4	target+non target+distance + IRL	0.30	3.62	0.19	0.38	3.77	0.27
5	lead vehicle+lane + IRL	0.30	3.57	0.23	0.29	3.51	0.28
6	target+non target + lane+lead vehicle + IRL	0.36	3.53	0.21	0.41	3.47	0.32
7	target+non target+distance + lane+lead vehicle + IRL	0.33	3.43	0.26	0.35	3.07	0.34
8	target+non target+distance + lane+driving task + IRL	0.51	3.41	0.31	0.57	2.18	0.59
9	target+non target+distance + lead vehicle+driving task + IRL	0.66	2.91	0.49	0.73	1.07	0.66
10	target+non target+distance+lane+lead vehicle+driving task + IRL	0.70	2.78	0.60	0.87	**0.87**	0.75
11	**MEDIRL**: target+non target+distance+lane+lead vehicle+driving task + speed + IRL	**0.74**	**2.51**	**0.61**	**0.89**	**0.88**	**0.78**

Table 4: Quantitative evaluation of the ablated versions of MEDIRL and full MEDIRL. All models trained on BDD-A train set and tested on EyeCar and BDD-A test sets. We mask out one part by setting the map(s) to zeros at each time.

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_β (β^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

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] Julia Beitner, Jason Helbing, Dejan Draschkow, and Melissa L-H Võ. Get your guidance going: Investigating the activation of spatial priors for efficient search in virtual reality. *Brain Sciences*, 11(1):44, 2021.

[3] Ali Borji, Majid Nili Ahmadabadi, Babak Nadjar Araabi, and Mandana Hamidi. Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, 28(7):1130–1145, 2010.

[4] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015.

[5] Ali Borji, Dicky N Sihite, and Laurent Itti. Probabilistic learning of task-specific visual attention. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 470–477. IEEE, 2012.

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[7] Anthony J-W Chen, Michael Britton, Gary R Turner, Jason Vytlacil, Todd W Thompson, and Mark D'Esposito. Goal-directed attention alters the tuning of object-based representations in extrastriate cortex. *Frontiers in human neuroscience*, 6:187, 2012.

[8] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2722–2730, 2015.

[9] Jin Chen, Huihui Song, Kaihua Zhang, Bo Liu, and Qingshan Liu. Video saliency prediction using enhanced spatiotemporal alignment network. *Pattern Recognition*, 109:107615, 2021.

[10] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9329–9338, 2019.

[11] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[12] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2146–2154, 2019.

[13] Tao Deng, Kaifu Yang, Yongjie Li, and Hongmei Yan. Where does the driver look? top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):2051–2062, 2016.

[14] Quoc Huy Do, Hossein Tehrani, Seiichi Mita, Masumi Egawa, Kenji Muto, and Keisuke Yoneda. Human drivers based active-passive model for automated lane change. *IEEE Intelligent Transportation Systems Magazine*, 9(1):42–56, 2017.

[15] Wolfgang Einhäuser, Charlotte Atzert, and Antje Nuthmann. Fixation durations in natural scene viewing are guided by peripheral scene content. *Journal of vision*, 20(4):15–15, 2020.

[16] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8554–8564, 2019.

[17] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. Dada-2000: Can driving accident be predicted by driver attention analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 4303–4309. IEEE, 2019.

[18] Keren Fu, Chen Gong, Irene Yu-Hua Gu, and Jie Yang. Normalized cut-based saliency detection by adaptive multi-level region merging. *IEEE Transactions on Image Processing*, 24(12):5671–5683, 2015.

[19] Mingfei Gao, Ashish Tawari, and Sujitha Martin. Goal-oriented object importance estimation in on-road driving videos. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5509–5515. IEEE, 2019.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[21] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3838, 2019.

[22] Chen Gong, Dacheng Tao, Wei Liu, Stephen J Maybank, Meng Fang, Keren Fu, and Jie Yang. Saliency propagation from simple to difficult. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2531–2539, 2015.

[23] Max Guangyu Li, Bo Jiang, Zhengping Che, Xuefeng Shi, Mengyao Liu, Yiping Meng, Jieping Ye, and Yan Liu. Dbus: Human driving behavior understanding system. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[24] Jason Helbing, Dejan Draschkow, and Melissa L-H Võ. Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition*, 196:104147, 2020.

[25] Zhiming Hu, Sheng Li, Congyi Zhang, Kangrui Yi, Guoping Wang, and Dinesh Manocha. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE transactions on visualization and computer graphics*, 26(5):1902–1911, 2020.

[26] Lai Jiang, Mai Xu, and Zulin Wang. Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm. *arXiv preprint arXiv:1709.06316*, 2017.

[27] Fumi Katsuki and Christos Constantinidis. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521, 2014.

[28] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pages 2942–2950, 2017.

[29] Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2020.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017.

[32] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798, 2017.

[33] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, and In So Kweon. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1947–1955, 2017.

[34] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015.

[35] Congcong Liu, Yuying Chen, Lei Tai, Haoyang Ye, Ming Liu, and Bertram E Shi. A gaze model improves autonomous driving. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–5, 2019.

[36] Feng Lyu, Nan Cheng, Hongzi Zhu, Haibo Zhou, Wenchao Xu, Minglu Li, and Xuemin Shen. Towards rear-end collision avoidance: adaptive beaconing for connected vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[37] Stefan Mathe and Cristian Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in neural information processing systems*, pages 1923–1931, 2013.

[38] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1408–1424, 2014.

[39] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2394–2403, 2019.

[40] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

[41] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[42] Eshed Ohn-Bar, Aditya Prakash, Aseem Behl, Kashyap Chitta, and Andreas Geiger. Learning situational driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11296–11305, 2020.

[43] Erfan Pakdamanian, Shili Sheng, Sonia Baee, Seongkook Heo, Sarit Kraus, and Lu Feng. Deeptake: Prediction of driver takeover behavior using multimodal data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[44] Anwesan Pal, Sayan Mondal, and Henrik I Christensen. "looking at the right stuff"-guided semantic-gaze for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11883–11892, 2020.

[45] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018.

[46] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[47] Constantin A Rothkopf, Dana H Ballard, and Mary M Hayhoe. Task and context determine where you look. *Journal of vision*, 7(14):16–16, 2007.

[48] Nathan Sprague and Dana Ballard. Eye movements for reward maximization. In *Advances in neural information processing systems*, pages 1467–1474, 2004.

[49] Hrvoje Stojić, Jacob L Orquin, Peter Dayan, Raymond J Dolan, and Maarten Speekenbrink. Uncertainty in learning, choice, and visual fixation. *Proceedings of the National Academy of Sciences*, 117(6):3291–3300, 2020.

[50] Benjamin W Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5–5, 2011.

[51] Ashish Tawari and Byeongkeun Kang. A computational framework for driver's visual attention using a fully convolutional architecture. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 887–894. IEEE, 2017.

[52] Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, and Adish Singla. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. In *Advances in Neural Information Processing Systems*, pages 4145–4155, 2019.

[53] Sabine Kastner Ungerleider and Leslie G. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341, 2000.

[54] Melissa Le-Hoa Võ, Sage EP Boettcher, and Dejan Draschkow. Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Current opinion in psychology*, 29:205–210, 2019.

[55] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[56] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017.

[57] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4903, 2018.

[58] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015.

[59] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3064–3074, 2019.

[60] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

[61] Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, Teresa Canas-Bajo, and David Whitney. Periphery-fovea multi-resolution driving model guided by human attention. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1767–1775, 2020.

[62] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian conference on computer vision*, pages 658–674. Springer, 2018.

[63] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2174–2182, 2017.

[64] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5188–5197, 2019.

[65] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 193–202, 2020.

[66] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020.

[67] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J Zelinsky, and Tamara L Berg. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746, 2013.

[68] Sara Zahedian, Kaveh Farokhi Sadabadi, and Amir Nohekhan. Localization of autonomous vehicles: Proof of concept for a computer vision approach. In *2019 ITS America Annual Meeting*. ITSWC, 2019.

[69] Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[70] Gregory J Zelinsky, Yupei Chen, Seoyoung Ahn, Hossein Adeli, Zhibo Yang, Lihan Huang, Dimitrios Samaras, and Minh Hoai. Predicting goal-directed attention control using inverse-reinforcement learning. *arXiv preprint arXiv:2001.11921*, 2020.

[71] Ruohan Zhang, Zhuode Liu, Luxin Zhang, Jake A Whritner, Karl S Muller, Mary M Hayhoe, and Dana H Ballard. Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the european conference on computer vision (eccv)*, pages 663–679, 2018.

[72] Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On learning intrinsic rewards for policy gradient methods. In *Advances in Neural Information Processing Systems*, pages 4644–4654, 2018.

[73] Sheng-hua Zhong, Yan Liu, Feifei Ren, Jinghuan Zhang, and Tongwei Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *Twenty-seventh AAAI Conference on Artificial Intelligence*, 2013.

[74] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.