TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
1–13

DOI: 10.1177/03611981221076121
journals.sagepub.com/home/trr
SAGE

Toward Explainable Artificial Intelligence for Early Anticipation of Traffic Accidents

Muhammad Monjurul Karim[1], Yu Li[1], and Ruwen Qin[1]

Abstract

Traffic accident anticipation is a vital function of Automated Driving Systems (ADS) in providing a safety-guaranteed driving experience. An accident anticipation model aims to predict accidents promptly and accurately before they occur. Existing Artificial Intelligence (AI) models of accident anticipation lack a human-interpretable explanation of their decision making. Although these models perform well, they remain a black-box to the ADS users who find it to difficult to trust them. To this end, this paper presents a gated recurrent unit (GRU) network that learns spatio-temporal relational features for the early anticipation of traffic accidents from dashcam video data. A post-hoc attention mechanism named Grad-CAM (Gradient-weighted Class Activation Map) is integrated into the network to generate saliency maps as the visual explanation of the accident anticipation decision. An eye tracker captures human eye fixation points for generating human attention maps. The explainability of network-generated saliency maps is evaluated in comparison to human attention maps. Qualitative and quantitative results on a public crash data set confirm that the proposed explainable network can anticipate an accident on average 4.57 s before it occurs, with 94.02% average precision. Various post-hoc attention-based XAI methods are then evaluated and compared. This confirms that the Grad-CAM chosen by this study can generate high-quality, human-interpretable saliency maps (with 1.23 Normalized Scanpath Saliency) for explaining the crash anticipation decision. Importantly, results confirm that the proposed AI model, with a human-inspired design, can outperform humans in accident anticipation.

Keyword

explainable artificial intelligence (XAI), automated driving system (ADS), saliency map

Autonomous driving research is advancing rapidly. Deep learning and computer vision are contributing technologies in this exciting journey (*1–3*). While autonomous vehicles are blooming, there are still cases where autonomous vehicles are involved in crashes (*4, 5*). Accident anticipation and avoidance are essential safety functions required for autonomous vehicles. From the dashboard camera (dashcam) video, an accident anticipation model aims to predict if an accident would occur shortly. Specifically, the model classifies the driving scene of the near future as one with or without accident risk. This accident anticipation model is a desired safety-enhancement capability not only for autonomous vehicles but also for countless human-driving vehicles. Successful anticipation of accidents from the widely deployed dashcams even just a few seconds ahead would effectively increase the situational awareness of human drivers, Advanced Driver Assistance Systems (ADAS), and autonomous vehicles to trigger a higher level of preparedness for accidents prevention.

Multiple computer vision-based deep learning models for the early anticipation of traffic accidents have been developed recently with outstanding performance (*6–9*). Each of the models may have millions of abstract parameters to learn from big data. Those learnable parameters are not directly attached to the physical nature of the problem to be solved. Therefore, despite the outstanding performance, the models' complex, black-box nature discourages societal acceptance. This issue is more acute for accident prediction models than many other Artificial Intelligence (AI) models. Accident anticipation is a high-stake, safety-critical function directly related to human lives. Thus, making AI models' decisions

[1]Department of Civil Engineering, Stony Brook University, Stony Brook, NY

Corresponding Author:
Ruwen Qin, ruwen.qin@stonybrook.edu

explainable to users who put their lives in the hands of AI is indispensable.

Importantly, human trust in autonomous driving technologies is a prerequisite for the mass adoption of autonomous vehicles (*10–15*). Shariff et al. (*11*) found that 78% of Americans reported fear of riding in an autonomous vehicle, and only 19% indicated that they would trust autonomous vehicles. If people cannot verify the safety and reliability of autonomous driving technologies, the perceived trustworthiness of the technologies is compromised (*12*). Shariff et al. also concluded that further research will need to identify the information required to form trustable mental models of autonomous vehicles. Ha et al. (*16*) experimentally confirmed that explanations increase human trust in autonomous driving. Researchers have reached a consensus that it is necessary to explain AI decisions for humans to trust AI systems. The European Union has already enacted a legal regulation allowing people to request an explanation of AI decisions if they are significantly affected by those decisions (*17*). Besides, the ability to explain the decision of an accident anticipation model would support insurance, legal, and regulatory entities to assess the model for the liability purpose (*18*). It is, therefore, imperative to make it transparent to people why AI models can anticipate traffic accidents. Achieving this goal will help remove the obstacle of integrating AI-powered accident anticipation into people's daily lives.

Explainable AI (XAI) is a rapidly growing research topic that aims to develop algorithms and tools to generate high-quality, interpretable, intuitive, and human-understandable explanations of AI models' decisions. Saliency maps are among the XAI tools that visually explain the decision made by a computer vision-based deep neural network. Salient regions in an image are those firmly relevant to the AI model's decision. Various methods exist for creating high-quality, easily interpretable saliency maps (*19–27*). While these models have built a methodological foundation for XAI, their assessment is challenging. Questions remain open on the trustworthiness of the explanation generated by XAI. One possible method to assess the quality of an XAI tool or algorithm is to let humans provide additional annotation and then evaluate the match level of human annotation with the explanation produced by saliency maps. However, human annotation can be expensive to acquire. XAI has been receiving growing attention in autonomous driving. Attempts have been made to explain the functions of various AI models for autonomous driving (*28–33*). Yet, XAI studies have not kept pace with the accelerating pace of AI-powered accident anticipation research.

This paper aims to develop an explainable deep neural network for early accident anticipation. The network comprises two major components. The first is a gated recurrent unit (GRU) network that analyzes the video captured by the dashcam to determine if an accident may occur shortly. This network mimics the human visual perception of accident risks during driving. Like the eyes of a driver, the dashcam captures the driving scene that contains comprehensive information about the surroundings. Similar to the brain of the driver, the deep neural network processes complex visual information to perceive the accident risk. The second component is the Gradient-weighted Class Activation Map (Grad-CAM) method that generates saliency maps to explain decisions made by the accident anticipation network. The saliency maps highlight pixels in a video that causally affect decisions made by the AI model. The explanation generated will help lift the psychological barrier humans have to enjoying the benefits promised by the AI-powered accident anticipation network. This study compares where the AI focuses with where the drivers look in predicting a future accident. This comparison reveals the quality of the saliency maps as a visual explanation.

In summary, the contributions of this paper are threefold:

- development of a deep neural network that learns the spatio-temporal relationship among visual features embedded in dashcam videos to predict if a traffic accident will occur shortly;
- integration of the developed deep network with the Grad-CAM method and its variants to product high-qualify saliency maps for interpreting the network's prediction; and
- collection of human gaze points to assess the quality of the XAI methods, which are affordable and efficient.

The remainder of this paper is organized as follows. The next section summarizes the related literature. The proposed methodology for creating the explainable accident anticipation network is then delineated. After that, the implementation details and experimental evaluation of the proposed method are discussed, followed by qualitative and quantitative analysis of the results. Finally, the conclusion and future work are summarized.

Literature Review

Saliency Maps as XAI Tools

Saliency maps can explain the decision of a neural network by highlighting regions of input images where the network responds the most in relation to its decision. Creating saliency maps for computer vision-based deep networks has become an XAI approach. Simonyan et al. (*19*) introduced a gradient-based method to generate saliency maps for Convolutional Neural Networks (CNN).

Their approach visualizes regions relevant to a CNN's decision after one forward and one backward propagation. The gradients from the end of the network are backpropagated and projected onto input images. Still, the vanilla gradient calculation produces noisy saliency maps. Therefore, subsequent methods, such as Guided Backpropagation (*20*), Deep Taylor (*21*), and Layerwise Relevance Propagation (LRP) (*22*), were developed to create better saliency maps by modifying the backpropagation algorithm.

Another line of research developed a method to create Class Activation Maps (CAM) that smooth saliency maps by localizing the class-specific regions in images (*23*). CAM requires removing the fully connected layers from the top of a trained CNN to put a global average pooling (GAP) layer followed by a single fully connected layer. The saliency map is then computed by summing up the activations of the last convolutional layer for individual output classes. However, a modification of the original architecture requires retraining the network. To overcome this drawback, Gradient-weighted CAM (Grad-CAM) generates high-quality saliency maps on the original network architecture without any modification (*24*). Recent extensions of Grad-CAM such as Grad-CAM++ (*25*), XGrad-CAM (*26*), and Eigen-CAM (*27*) are powerful XAI tools for visualizing the class-specific decision that deep neural networks make.

XAI for Autonomous Driving

XAI is receiving growing attention in autonomous driving. For example, Bojarski et al. (*28*) proposed an activation-based method to backpropagate activations for obtaining smooth heat maps. This method for explaining CNN works in a real-time manner to be integrated by autonomous driving. Kim et al. (*29*) adopted an attention-based method to filter out non-salient image regions and display only those causally affecting the steering control of a stand-alone vehicle. Similarly, Cultrera et al. (*30*) also used an attention model to visualize the perception of deep networks for autonomous driving. Autonomous driving has employed saliency to explain AI models for navigation, lane change detection, and driving behavior reasoning (e.g., hazard stop or red light stop) (*33*). All these methods achieved an impressive performance in explaining decisions that AI models made for critical tasks of autonomous driving. However, the explanation of deep networks for early anticipation of traffic accidents is less common. One XAI study partially related to accident anticipation is noticed, which attempts to explain object-induced actions of vehicles (*34*). This study focuses on scene understanding, highlighting salient objects in the input images which potentially lead to a hazard. The network architecture selects potential objects from the region proposals proposed by Faster R-CNN (*35*) without considering spatio-temporal relational information. It, therefore, disregards motion information that is an essential determinant of accidents.

Methodology

Figure 1 illustrates the proposed method for creating the explainable accident anticipation network. The video captured by the dashcam, as a sequence of frames indexed by t, flows into a feature extractor. The feature extractor extracts a feature map $\boldsymbol{A}_t$ from frame t. After passing a dense layer, the feature map becomes a feature

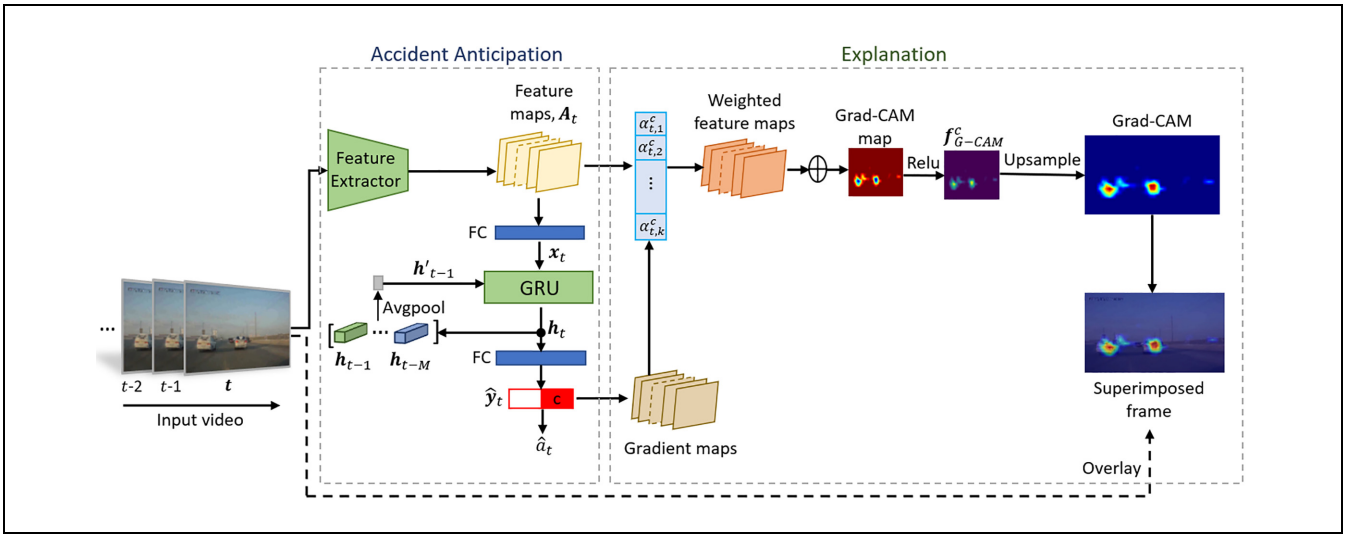

Figure 1. Overview of the proposed method to create the explainable accident anticipation network.
Note: GRU = gated recurrent unit.

vector, $\boldsymbol{x}_t$, which goes to a GRU to learn the hidden representation of the frame, $\boldsymbol{h}_t$. Using the hidden representation, the network predicts the scores of the accident class c and non-accident class $!c$, denoted by $\hat{\boldsymbol{y}}_t = [\hat{y}_t^c, \hat{y}_t^{!c}]$. Gradient maps are calculated by backpropagating the prediction score with respect to the feature maps obtained from the feature extractor. Then, importance weights $\boldsymbol{\alpha}^c$ are calculated from the gradients to aggregate all channels of the feature map into the Grad-CAM map that is further used to compute the final Grad-CAM saliency maps. Details of the method are discussed below.

Feature Extraction

The base feature extractor used in this study is the ResNet50 network, initialized using parameters pretrained on ImageNet (*36*). The feature extractor extracts a feature map $\boldsymbol{A}_t \in \mathbb{R}^{K \times U \times V}$ from frame t, for any t. That is, the feature map has K channels, and the height and width of each channel are U and V, respectively. Then, the flattened feature maps are passed through a dense layer to become a d dimensional feature vector, $\boldsymbol{x}_t \in \mathbb{R}^d$.

Spatio-temporal Relational Learning with GRU

A Recurrent Neural Network (RNN) is a powerful tool for spatio-temporal sequential learning. Spatio-temporal relationships among image features contain important cues for accident anticipation. This study uses GRU, a particular type of RNN, to learn spatio-temporal relationships among the features by updating the hidden representation of each frame, $\boldsymbol{h}_t$. GRU has two gates, a reset gate $\boldsymbol{g}_t^{(r)}$ and an update gate $\boldsymbol{g}_t^{(u)}$, which retain the most relevant information from the video sequence by filtering out irrelevant information. The data flowing through the GRU are expressed mathematically in Equations 1–4:

$$\boldsymbol{g}_t^{(r)} = \sigma(\boldsymbol{W}_g^{(r)}\boldsymbol{x}_t + \boldsymbol{B}_g^{(r)}\boldsymbol{h}'_{t-1}), \tag{1}$$

$$\boldsymbol{r}_t = \tanh(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{B}_r(\boldsymbol{g}_t^{(r)} \circ \boldsymbol{h}'_{t-1})), \tag{2}$$

$$\boldsymbol{g}_t^{(u)} = \sigma(\boldsymbol{W}_g^{(u)}\boldsymbol{x}_t + \boldsymbol{B}_g^{(u)}\boldsymbol{h}'_{t-1}), \tag{3}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{g}_t^{(u)}) \circ \boldsymbol{r}_t + \boldsymbol{g}_t^{(u)} \circ \boldsymbol{h}'_{t-1}, \tag{4}$$

where

σ represents the sigmoid activation,

$\circ$ is the element-wise product operator,

$\boldsymbol{W}$'s and $\boldsymbol{B}$'s ($\in \mathbb{R}^{d \times d}$) are learnable parameters, and

$\boldsymbol{h}'_{t-1}$ is the average pooled hidden representations of the past M frames:

$$\boldsymbol{h}'_{t-1} = \text{avgpool}([\boldsymbol{h}_{t-1}, \boldsymbol{h}_{t-2}, \ldots, \boldsymbol{h}_{t-M}]). \tag{5}$$

This study found that the average pooling of the most recent M frames' hidden representations is better than the hidden representation of the last single frame in learning the contextual information.

The hidden state $\boldsymbol{h}_t$ is then projected onto the prediction scores of the two classes—accident and no-accident—using a fully connected layer, f_c, with parameters $\boldsymbol{W}_0$ and $\boldsymbol{B}_0$ ($\in \mathbb{R}^{d \times d}$):

$$\hat{\boldsymbol{y}}_t = [\hat{y}_t^c, \hat{y}_t^{!c}] = f_c(\boldsymbol{h}_t; \boldsymbol{W}_0, \boldsymbol{B}_0). \tag{6}$$

After that, the probability of seeing an accident shortly, predicted at time t, is obtained by the softmax operation on $\hat{\boldsymbol{y}}_t$:

$$\hat{a}_t = \exp(\hat{y}^c)/[\exp(\hat{y}^c) + \exp(\hat{y}^{!c})] \tag{7}$$

Generating Saliency Maps with Grad-CAM

This study uses Grad-CAM for generating saliency maps for the prediction of accident class c. Input video frames indexed by t are fed to the network to calculate the Grad-CAM saliency maps for class c. Let y_t^c be the score for class c predicted at frame t, calculated in Equation 6. First, the gradients of y_t^c on the k th channel of the feature map $\boldsymbol{A}_{t,k}$ are computed for all channels. Then, the gradients are averaged along the width and height of each channel to obtain the channel-wise importance weights $\alpha_{t,k}^c$:

$$\alpha_{t,k}^c = \frac{1}{UV}\sum_{u=1}^{U}\sum_{v=1}^{V}\frac{\partial y_t^c}{\partial \boldsymbol{A}_{t,k}(u,v)}. \tag{8}$$

Afterwards, the importance weights are used to aggregate the K channels of the feature map to get the Grad-CAM map. Finally, the activation function ReLU is applied to find the rectified Grad-CAM map for class c:

$$\boldsymbol{f}_{\text{G-CAM}}^c = ReLU\left(\sum_{k=1}^{K}\alpha_{t,k}^c\boldsymbol{A}_{t,k}\right). \tag{9}$$

The Grad-CAM map $\boldsymbol{f}_{\text{G-CAM}}^c$ has the same spatial dimension as $\boldsymbol{A}_{t,k}$. That is, $\boldsymbol{f}_{\text{G-CAM}}^c \in \mathbb{R}^{U \times V}$. Its resolution is lower than that of the original video frame. Therefore, $\boldsymbol{f}_{\text{G-CAM}}^c$ is up-sampled to have the same size of the input frame via the bilinear interpolation. The resized Grad-CAM map is superimposed with the input frame to explain the network's decision visually.

Implementation and Experimental Evaluation

The Data Set

A publicly available data set called the Car Crash Dataset (CCD) (*6*) is used to train and evaluate the proposed accident anticipation network. CCD comprises

videos that were captured by dashcams mounted on vehicles. The positive class of the data set contains 1,500 videos that each contain an accident. The negative class has 3,000 videos without any accident. The data set has diverse driving environment attributes. All the videos in CCD are trimmed to 5 s, and each video has 50 frames (i.e., 10 frames per second [fps]). The data set is split into the training data set with 3,600 videos and the testing data set with 900 videos.

Model Training

The training data set, which comprises N ($=3{,}600$) videos indexed by n, is used to train the network. The video-level label is $l_n = 1$ if video n is positive and $l_n = 0$ if it is negative. Negative videos use the vanilla cross entropy loss function. Positive videos use an exponential cross entropy loss function to encourage early anticipation of accidents. That is, the loss for a positive video is near zero when time is far away from the accident, increases gradually as time is approaching the accident occurrence, and reaches the same level for negative videos at the accident occurrence and onward. Finally, the total loss on the training data set is:

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^{N}\Bigg[&-(1-l_n)\sum_{t=1}^{T}\log(1-\hat{a}_{t,n}) \\ &-l_n\sum_{t=1}^{T}\exp\left[-\max\left(\tfrac{\tau-t}{f},0\right)\right]\log(\hat{a}_{t,n})\Bigg].\end{aligned} \tag{10}$$

where $\hat{a}_{t,n}$ is the softmax probability that video n belongs to the accident class, predicted at frame t.

The proposed network was trained and tested using an Nvidia V100 GPU with 32GB of memory. Input video frames were down sampled to 224×224 before being fed to the network. The dimension of the feature map $\boldsymbol{A}_t$ at the last convolution layer of the feature extractor is $14(U) \times 14(V)$ with 512 (K) channels. The dimension of the feature vector before going to the GRU is 2,048 (d). The dimension of hidden representations output from the GRU is 256. The number of hidden representations for average pooling is 3 (M). The network was trained with the learning rate 0.0001, the batch size 10, and ReduceLROnPlateu as the learning rate scheduler. Adam optimizer was used to optimize the network for 30 epochs.

Evaluation Metrics for the Accident Anticipation Model

Assessment of the accident anticipation network was performed to determine how precisely and early the network can anticipate traffic accidents. Two metrics described below are used for the assessment.

Average Precision. On a testing video, if the softmax probability of accident, $\hat{a}_t$, exceeds a pre-specified threshold value $\bar{a}$ before an accident occurs, the video is predicted as positive. Otherwise, the prediction is negative. Accordingly, the recall (R) and precision (P) for the model were calculated based on the testing data set:

$$\mathrm{R} = \frac{\#\,\text{true positive predictions}}{\#\,\text{positive videos}}, \tag{11}$$

$$\mathrm{P} = \frac{\#\,\text{true positive predictions}}{\#\,\text{positive predictions}}. \tag{12}$$

Recall and precision values depend on the choice of classification threshold. A precision-recall curve can be created given various threshold values, and average precision (AP) is the area below this curve:

$$\mathrm{AP} = \int \mathrm{P_R}\, d\,\mathrm{R}. \tag{13}$$

where $\mathrm{P_R}$ is the precision at a given recall value. AP measures how accurate the network is in anticipating accidents.

For a real-world implementation, the user may require a relatively high recall value. Therefore, precision at the recall value 80%, denoted by $P_{80\,\mathrm{R}}$, was also calculated in this study.

Time-to-Accident. A positive video where the accident occurs at τ is predicted as positive when the probability $\hat{a}_t$ exceeds the classification threshold for the first time. The time-to-accident (TTA) is the time from the prediction to the accident occurrence:

$$\mathrm{TTA}(\bar{a}) = \max\{\tau - t | \hat{a}_t > \bar{a}, 0 \leqslant t \leqslant \tau\}. \tag{14}$$

TTA measures how early the network can predict an accident, which is subject to the choice of the threshold value. The expected value of $\mathrm{TTA}(\bar{a})$ is the mean TTA:

$$\mathrm{mTTA} = \mathrm{E}_{\bar{a}}[\mathrm{TTA}(\bar{a})]. \tag{15}$$

This study also calculates $\mathrm{TTA}_{\,80\,\mathrm{R}}$ in correspondence to $\mathrm{P}_{\,80\,R}$.

Creating Human Attention Maps as a Reference

Drivers' visual attention is influenced by important visual cues in the traffic scene, such as a pedestrian, a cyclist, changes of traffic lights, or abnormal behavior of other vehicles. Therefore, drivers' gaze behavior can be used as the proxy for their attention. In this study, to evaluate the explainability of the developed network, human attention maps are computed and compared with the saliency maps generated by Grad-CAM.

To obtain human attention maps, an eye tracking experiment was designed and performed using a Tobii Pro Fusion (*37*) eye tracker. Tobii Pro Fusion is a screen-based eye tracker. Twelve volunteers participated in the experiment including three females and nine males. The age of the participants ranged from 20 to 43. Their driving experience ranged from 3 months to 18 years. All the participants passed the eye tracker calibration test and are thus qualified for the eye tracking experiment.

The experiment chose 100 videos randomly from the test data set of CCD to create an XAI test data set. Fifty of the videos are positive videos, and the rest are negative videos. To assure the diversity of the scenes, the selected video clips contain a proportional number of environmental attributes. For example, 67% of the video clips are normal weather and the remaining 33% are snowy and rainy weather. These videos are in a random sequence so that participants do not know the class of the next video to watch. Furthermore, all the video clips are 5 s in length. For any positive video, the crash starting time is randomly placed in the last 2 s of the video clip. Given the randomness, participants cannot predict the timing of crash occurrence from their experiences of watching other videos but the visual cues of the accident risk. Participants assume they are drivers when watching these videos. The eye tracker captures the participants' gaze data, including the timestamp and coordinates of each gaze point on the video frames. Each participant performed the experiment once. In total, the experiment recorded about 720,000 gaze points, approximately 144 gaze points per frame. Gaze points could be classified into fixation, saccade, and unknown based on the angular speed of drivers' gaze movement.

A Gaussian filter of 30×30 pixels was used to convolute the count of gaze points on each frame into the human attention map. Figure 2 shows a few samples of such human attention maps. The green dots in the first row of the figure are human gaze points. The second row shows the attention maps created after applying the Gaussian filter.

While fixation points are usually used to create human attention maps, either licensed software is required or a complex algorithm needs to be developed, to extract fixation points from gaze points. This study found that fixation points are the dominant class for drivers, accounting for 93% of the collected gaze points. Therefore, it introduced a simplified method that uses all the gaze points to create human attention maps. While this approximation introduces a small amount of noise to the attention maps, objects that drivers attend to still stand out in the maps.

Obtained human attention maps are passed through a step filter to create fixation maps of binary pixel values. Pixels of value 1 in a fixation map are considered as fixated pixels.

Evaluation Metrics for XAI

This study evaluated the Grad-CAM method's ability to explain the prediction by the accident anticipation network by comparing the generated saliency maps to human fixation maps. The study considered both location-based and distribution-based metrics delineated below.

Normalized Scanpath Saliency. Normalized Scanpath Saliency (NSS) measures the correspondence between saliency maps of the accident anticipation network and human fixation maps. Let $\hat{\boldsymbol{S}}$ be the Grad-CAM saliency map of a frame and $\boldsymbol{F}$ be the fixation map of the same frame.

$$\mathrm{NSS}(\hat{\boldsymbol{S}}, \boldsymbol{F}) = \frac{1}{\sum_i F_i} \sum_i \frac{\hat{S}_i - \mu_{\hat{S}}}{\sigma_{\hat{S}}} F_i \tag{16}$$

where

i is the pixel index of $\hat{\boldsymbol{S}}$ and $\boldsymbol{F}$,
$\hat{S}_i$ is the value of $\hat{\boldsymbol{S}}$ at pixel i, and
F_i is the value of $\boldsymbol{F}$ at pixel i.

Mathematically, NSS is the average of normalized values at pixels of the saliency map where human fixations fall

Figure 2. Human attention maps obtained from recorded gaze points.

Table 1. Accident Anticipation Performance on CCD

Experiment ID	*M* (#)	AP (%)	mTTA (s)	P_{80R} (%)	TTA_{80R} (s)
1	1	93.77	4.45	92.31	4.32
2	3	94.02	4.57	93.02	4.50

Note: CCD = car crash data set; AP = average precision; TTA = time-to-accident; mTTA = mean time-to-accident.

on. A positive NSS value means at a positive chance that the AI-rendered saliency map and the human fixation map have correspondence. The larger the NSS value, the stronger the correspondence.

Area Under Receiver Operating Characteristic (ROC) Curve. The saliency map of a video frame can be seen as the prediction of the human fixation map of the frame. After applying a specific threshold value to the saliency map, the false positive rate and true positive rate of the binary classification result by the saliency map are obtained. The Receiver Operating Characteristic (ROC) curve can be created to measure the trade-off between true and false positives at various thresholds. The Area Under the Curve (AUC) is then computed to measure how well the model performs. Two different variants of AUC, AUC-J (*38*) and AUC-B (*38*), are computed in this study. The range of AUC is [0, 1]. The larger the AUC value, the higher chance that the saliency map and the human fixation map have correspondence.

Kullback–Leibler Divergence. Kullback–Leibler (KL) divergence is generally used to estimate dissimilarity between two distributions. This study used KL to compare the network-generated saliency maps with human fixation maps. Given a saliency map $\hat{\boldsymbol{S}}$ and the corresponding human fixation map $\boldsymbol{F}$, KL divergence can be approximated as:

$$KL(\hat{\boldsymbol{S}}, \boldsymbol{F}) = \sum_i F_i \log\left(\epsilon + \frac{F_i}{\epsilon + \hat{S}_i}\right), \qquad (17)$$

where ϵ is a regularization constant. The range of KL divergence score is [0, ∞]. A low KL divergence score indicates a better approximation of the fixation map by the network-generated saliency map.

Results and Discussion

Performance of the Accident Anticipation Model

The accident anticipation model was assessed on the test data set of CCD. Two different experiments were conducted in this study by changing the number of hidden representations (M) fed to the GRU. During training, the network optimized the parameters by backpropagating the loss function. A set of trade-off solutions between AP and mTTA were obtained from different epochs of the training process. This paper only reports the results when the highest AP is achieved; these are shown in Table 1.

In Table 1, the first experiment used the hidden representation of that last frame ($M = 1$) to update the hidden representation of the current frame, whereas the second experiment mean-pooled the hidden representations of the last three frames ($M = 3$). The result shows how the integration of several recent frames' hidden representations helps the network learn the temporal information better than if it were only using a single frame. The performance achieved in the second experiment confirms that the proposed model can predict accidents very earlier (mTTA = 4.57 s and TTA_{80R} = 4.50 s) with a very low false alarm rate (AP = 94.02% and P_{80R} = 93.02%).

Figure 3 further illustrates an example of accident anticipation by the proposed network. In the figure, sample frames of a video clip are shown on the top, the corresponding saliency maps are in the middle, and the time series of the accident anticipation probability $\hat{a}_t$ denoted by the red colored curve is at the bottom. In the video, the accident starts at frame #38. For illustration purposes, a threshold value of 0.5 ($\bar{a}$) is set to trigger the prediction of an accident in this example. The probability of accident anticipation reached the threshold at frame #7, which yields a TTA of 3.2 s. That is, the network predicted the accident 3.2 s before it happenned. Hot spots in the saliency maps describe the network's concentration in the accident anticipation. Very early in the video (e.g., frame #0), activations of the accident class were relatively sparse. Through learning the spatio-temporal relationships, the network concentrated more on objects that might be involved in or be affected by an accident. In this example, two vehicles, A and B, at a relatively long distance from the ego-vehicle, were first involved in a crash. Two additional vehicles, C and D, closer to the ego-vehicle were affected by the crash. The one on the left adjacent lane (vehicle C) failed to avoid the crash, whereas the one in front (vehicle D) successfully avoided the accident. The saliency maps show that the network started attending to vehicles *C* and *D* very earlier (see frames #10 and #30). At frame #38, the crash started

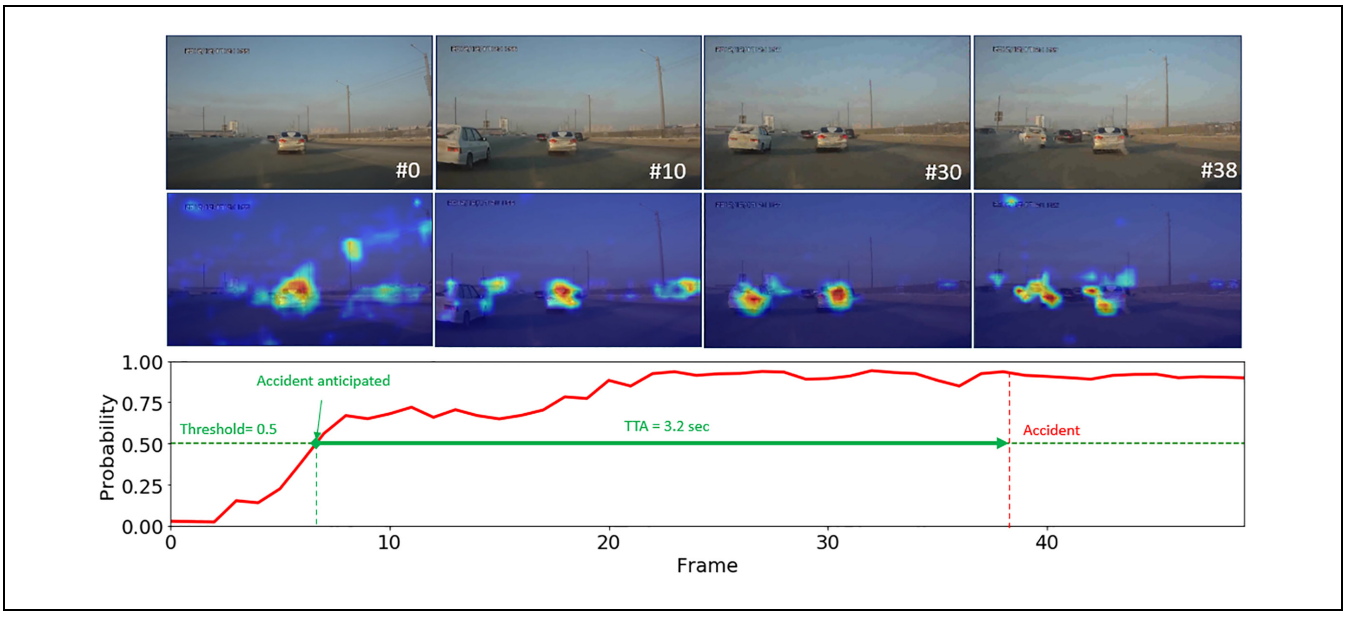

Figure 3. Visualizing the accident anticipation by the proposed network for explanation.
Note: TTA = time-to-accident.

with vehicles A and B that were involved first, and the saliency on vehicles C and D are split, and part of the saliency split apart goes toward vehicles A and B.

Performance of Explainable Artificial Intelligence (XAI) Methods

Qualitative Comparison. Figure 4 illustrates the qualitative results of several examples selected from the tested accident videos. The first column shows input video frames, the second column lists drivers' fixation maps. The third column is the human attention maps overlaid with the input image to show where drivers attend. The last four columns, from the left to the right, are the saliency maps generated by Grad-CAM, Grad-CAM++, XGrad-CAM, and Eigen-CAM, respectively. The saliency maps are also overlaid with their respective input images to conveniently explain the accident anticipation decision made by the proposed accident anticipation network. All the selected sample frames are within 2 s of the accident occuring, where objects involved in or affected by the accident are within the frames.

Humans have a natural tendency to focus on salient objects that have visual significance in understanding the situation. From the human attention maps (column 3 of Figure 4), it is evident that humans attend to vehicles or pedestrians that may be involved in or affected by a traffic accident. The hottest spots in the saliency maps created by the Grad-CAM method (in column 4) closely overlap with the locations where humans fixate. This indicates the proposed accident anticipation network predicts high saliency values on the traffic agents involved in or affected by the incident and, accordingly, derives the prediction successfully. Saliency maps created by XGrad-CAM (column 6) are very similar to those generated by Grad-CAM (column 4). Saliency maps obtained by Grad-CAM++ (Column 5) are sparser, where a lot of regions irrelevant to the accident are also considered as salient regions. Those are less capable of describing the network's decision. Similarly, Eigen-CAM creates very random saliency maps (column 7) that do not explain the network's decision well.

The comparative analysis based on Figure 4 reveals that the accident anticipation network developed in this study predicts a future accident by focusing on the most salient regions, just like a human does. The Grad-CAM method, as well as the XGrad-CAM method, can generate high-quality saliency maps to explain how the proposed accident anticipation network makes a decision. Since the saliency maps generated by Grad-CAM and XGrad-CAM are interpretable to humans, they help users of the accident anticipation network establish their faith and trust in the underlying AI model.

Quantitative Assessment. The study further compared saliency maps generated by the XAI methods with human fixation maps using the evaluation metrics NSS, AUC-J, AUC-B, and KL divergence. For computing the NSS, a filter with a threshold value of 0.1 was applied to human attention maps to obtain binary fixation maps. Results for both the positive and the negative classes are summarized in Table 2. It should be noted that an XAI

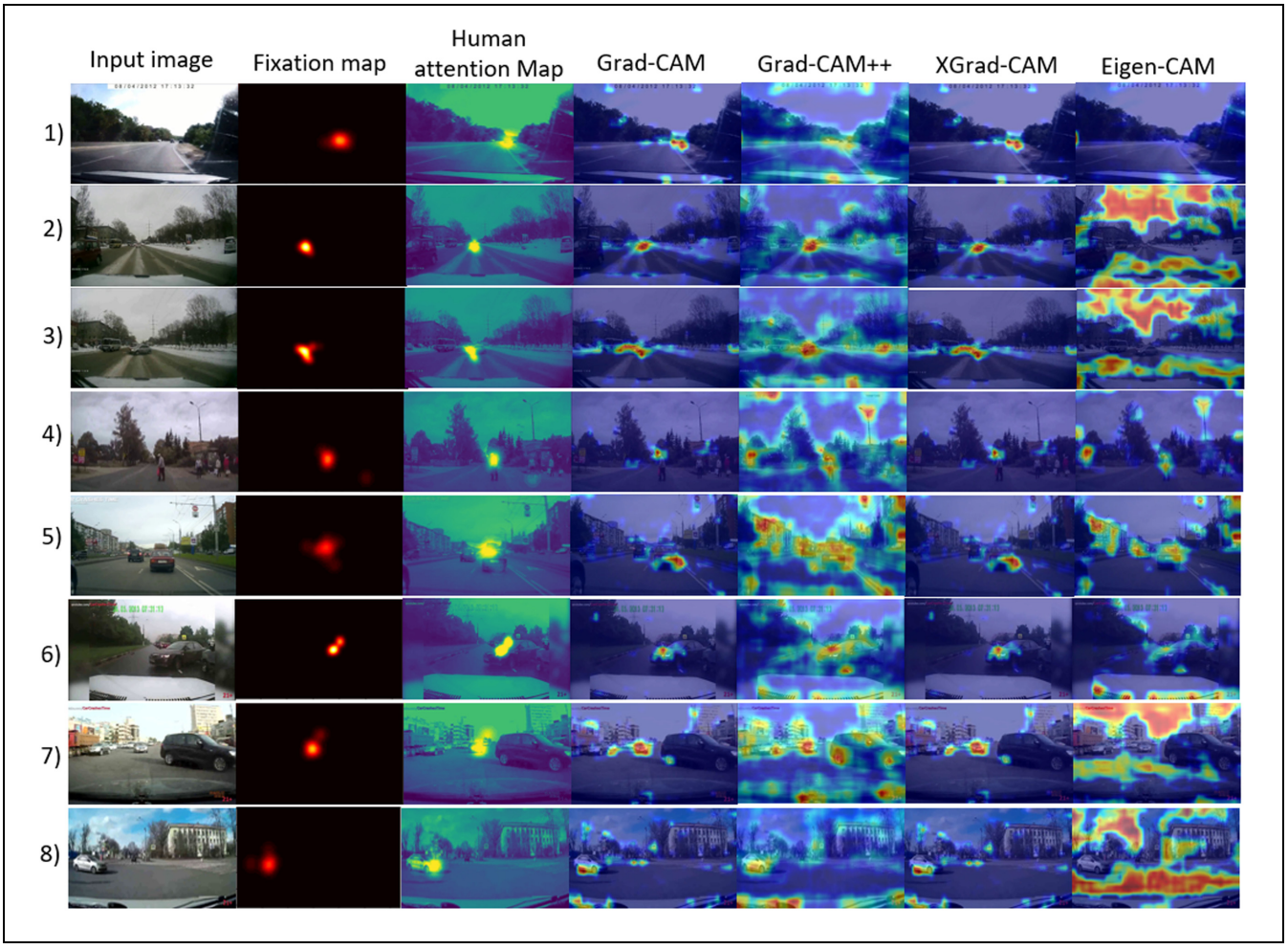

Figure 4. Visualizing the explanation generated by different explainable artificial intelligence methods.

Table 2. Performance Comparison of Explainable Artificial Intelligence (XAI) Methods

XAI methods	Positive				Negative			
	NSS	AUC-J	AUC-B	KL	NSS	AUC-J	AUC-B	KL
Grad-CAM	1.23	0.70	0.65	4.55	0.50	0.63	0.60	4.63
Grad-CAM++	1.13	0.76	0.75	12.75	0.86	0.73	0.72	8.02
XGrad-CAM	1.23	0.70	0.65	4.55	0.50	0.62	0.60	4.63
Eigen-CAM	0.10	0.50	0.49	13.13	0.04	0.50	0.48	13.24

Note: NSS = Normalized Scanpath Saliency; AUC-J = Area under Receiver Operating Characteristic Curve-Judd; AUC-B = Area under Receiver Operating Characteristic Curve-Borji; KL = Kullback-Leibler divergence.

method is more favorable if it has larger values of NSS and AUC and a smaller value of KL divergence.

It can be observed from Table 2 that Grad-CAM and XGrad-CAM achieve a similar performance in that they have near-identical values on all the metrics and on both classes. On the positive class, both of these two XAI methods receive reasonably good scores on all metrics. It should be mentioned that human fixation maps do not set an ideal performance benchmark for the XAI methods, but a reasonable reference. This is because humans attend to one spot at one time, and they may miss some critical salient regions in fast-changing driving scenes. Whereas the proposed AI network can attend to many salient regions or traffic agents in a video frame. Since human fixation maps, which are used as the ground truth for assessing the XAI method, are not perfect

saliency maps of accident risks, values of the metrics on the positive class were underestimated. The Grad-CAM++ method receives a lower NSS value (1.13) but higher AUC-J value (0.76) and AUC-B value (0.75) on the positive class than both Grad-CAM and XGrad-CAM. The higher AUC values that the Grad-CAM++ method achieved on the positive class are a result of the AUC metrics not penalizing false positives whereas NSS does. Therefore, Grad-CAM++ achieves better AUC values although it generated sparse saliency maps that are visually different than human attention maps on the positive classes. Eigen-CAM performs poorly on all four metrics on the positive class. The quantitative results suggest that the Grad-CAM method and the XGrad-CAM method can better explain the decision of the accident anticipation network by rendering high-quality saliency maps on the positive class.

The performance of the XAI methods on the negative class was also evaluated in this study, with results presented in the right portion of Table 2. Values of the metrics on the negative class differ from the values on the positive class. Using the metric NSS as an example, the NSS value of any XAI method on the negative class is smaller than that on the positive class. This is because, when a driving scene is normal and low in the accident risk, the saliency maps that XAI methods generate for frames of the negative class are sparsely dispersed among different traffic agents or regions in the frames. However, humans always have a certain degree of fixation, regardless of the class of the driving scene. When the driving scene is normal, drivers fixate on regions and traffic agents for the purposes of navigating and conforming with driving rules. When the driving scene is risky, their attention is partially attracted by the traffic agents involved in or affected by the accident. On videos of the negative class, regions with human eye fixations do not, therefore, have much saliency value for the accident risk. This explains the reason for obtaining low NSS values on the negative class.

The NSS value of any XAI method on the negative class is below 1 and lower than the value on the positive class. In particular, the between-class differences in the NSS values that Grad-CAM and XGrad-CAM obtain (1.23 versus 0.50) are the largest among the four methods. The AUC values of Grad-CAM, Grad-CAM++, and XGrad-CAM on the negative classes are also lower than the values on the positive classes. Again, the between-class differences in AUC values that Grad-CAM and XGrad-CAM achieve are greater than Grad-CAM++. Compared with the KL divergence values on the positive class, the KL values of Grad-CAM, XGrad-CAM, and Eigen-CAM on the negative classes increase a little, whereas the value of Grad-CAM++ decreases. The comparison shows that Grad-CAM and XGrad-CAM are better than Grad-CAM++ and Eigen-CAM in differentiating risky driving scenes from normal scenes. The effectiveness of the metrics, in descending order, is NSS, AUC, and KL.

Inference Speed

The inference speed of the proposed accident anticipation model is critical because accident anticipation needs to be real-time. This study thus evaluated the efficiency of the proposed method with regard to inference speed. By testing on the test data set, it is found that the proposed method can process a video at a speed of 8.5 fps from loading a video frame to generating the prediction score. This study further evaluated the inference speed for the saliency map generation process. On average, Grad-CAM, Grad-CAM++, and XGrad-CAM take 11 ms, 13 ms, and 12 ms, respectively, for producing the saliency map for each video frame. However, Eigen-CAM takes a longer time, about 1.3 s per frame.

Challenges and Opportunities

The output of the accident anticipation network developed in this paper is binary because the network classifies the driving scene of the near future as one with or without the accident risk. The current system can anticipate traffic accidents regardless of the accident type. Once sufficient training data become available, the current network could be be extended to become a multiclass network to differentiate the accident type if an accident is anticipated to occur shortly. According to the Traffic Safety Facts Annual Report provided by National Highway Traffic Safety Administration (NHTSA), accidents resulting in fatalities, injuries, or property damages are associated with various types of harmful events: (1) no collision with a motor vehicle in transport; (2) rear-end; (3) head-on; (4) angle; (5) sideswipe; and (6) unknown. In the period from 2015 to 2019, 163,350 (96.65%) of fatal crashes in the United States were of the first four types (*39*).

Indeed, a forward-facing dashcam will not capture all types of accident. For example, in a potential sideswipe accident, the vehicle being collided with cannot capture the risk from its dashcam. However, dashcams are a low-cost sensor type widely deployed in many vehicles. The field of view limitation can be addressed well if most vehicles are equipped with dashcams and are accessible to an accident anticipation system. For example, the vehicle likely to cause a sideswipe accident may see the risk from its dashcam. Other surrounding cars not involved in the accident may also perceive the sideswipe risk from their cameras. Likewise, in a potential rear-end accident, the car appraoching from behind can perceive the accident

risk from its dashcam and thus try to avoid it, although the vehicle in front does not. Furthermore, adding additional camera sensors can increase the field of view. As we are moving toward the era of Vehicle-to-Vehicle (V2V) communication, vehicles that have anticipated an accident can share the information with others to alert them to possible accidents. V2V communication will further broaden the impact of dashcam-based accident anticipation.

Conclusions

This paper presented an explainable deep neural network for early anticipation of traffic accidents from dashcam videos. The proposed network is a GRU that learns the spatio-temporal relationship between visual features of accidents by updating its hidden representations. Aggregating the hidden representations of multiple recent frames, the GRU learns the temporal-contextual information better, thus improving the network's performance. The experimental evaluation on the CCD data set confirms that the proposed network can anticipate accidents very early (with 4.57-s mTTA) and accurately (with 94.02% AP). Additionally, the Grad-CAM is integrated into the proposed accident anticipation network to produce saliency maps that explain the network's decision visually. Human gaze data on a dashcam video data set were collected to compare with the saliency maps generated by the explainable network. Four variants of XAI methods were further evaluated in this study. Out of these four methods, Grad-CAM and XGrad-CAM methods are most suitable because they generate a high-quality visual explanation for the accident anticipation decision made by the network. Qualitative and quantitative evaluations both confirm that the proposed accident anticipation network has reliable and visually interpretable performance. The proposed explainable network can, therefore, not only build drivers' confidence in reliable AI models but address some drivers' blind trust in unreliable AI models.

This paper has identified room for improvement. For example, the human subject experiment was performed in a laboratory setting. However, field testing scenarios can be more complex than in the controlled laboratory experiment, thus affecting human attention. Capturing human gaze data from actual driving conditions will help calibrate the results of evaluating the XAI methods. This study also found that the proposed AI method can surpass humans in finding salient regions to support the accident anticipation. Thus, to further strengthen people's trust in AI, an evaluation method will be developed to measure the strengths of AI over humans or vice versa. The accident anticipation network can enhance its performance further using human attention maps as a separate input channel.

Camera sensors have certain limitations as well as merits, just like every other sensor type. For example, cameras may be blocked by dirt or mud while driving. The limitations of cameras do not restrict them from making positive contributions to transportation safety and autonomous driving. Moreover, it is unlikely that autonomous vehicles can rely on a single system to sense the environment and navigate themselves. This study envisions sensor fusion as a solution to address the limitations of individual sensor types. Individual sensor technologies should be developed to maximize the effectiveness of sensor fusion. The advancement and synergy of these sensor technologies will positively contribute to the full driving automation.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: M. M. Karim, R. Qin; data collection: Y. Li; analysis and interpretation of results: M. M. Karim, R. Qin; draft manuscript preparation: M. M. Karim, R. Qin. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: All the authors received financial support from National Science Foundation through the award ECCS-2026357.

ORCID iDs

Muhammad Monjurul Karim https://orcid.org/0000-0002-7830-1407
Ruwen Qin https://orcid.org/0000-0003-2656-8705

Data Accessibility Statement

Data, models, and code that support the findings of this study are available at https://github.com/monjurulkarim/xai-accident.

References

1. Janai, J., F. Güney, A. Behl, and A. Geiger. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and Trends®in Computer Graphics and Vision*, Vol. 12, No. 1–3, 2020, pp. 1–308.
2. Eskandarian, A., C. Wu, and C. Sun. Research Advances and Challenges of Autonomous and Connected Ground

Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 22, No. 2, 2021, pp. 683–711.
3. Muhammad, K., A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de. Albuquerque. Deep Learning for Safe Autonomous Driving: Current Challenges and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 22, No. 7, 2021, pp. 4316–4336.
4. Wiggers, K. Waymo's Driverless Cars were Involved in 18 Accidents Over 20 Months, 2020. https://venturebeat.com/2020/10/30/waymos-driverless-cars-were-involved-in-18-accidents-over-20-month/. Accessed March, 2021.
5. California, DMV. Autonomous Vehicle Collision Reports, 2021. Accessed September, 2021. https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/.
6. Bao, W., Q. Yu, and Y. Kong. Uncertainty-Based Traffic Accident Anticipation With Spatio-Temporal Relational Learning. *Proc., 28th ACM International Conference on Multimedia*, Association for Computing, New York, NY, Seattle, WA, 2020, pp. 2682–2690.
7. You, T., and B. Han. Traffic Accident Benchmark for Causality Recognition. *Proc., European Conference on Computer Vision*, Springer, Glasgow, 2020, pp. 540–556.
8. Li, Y., M. M. Karim, R. Qin, Z. Sun, Z. Wang, and Z. Yin. Crash Report Data Analysis for Creating Scenario-Wise, Spatio-Temporal Attention Guidance to Support Computer Vision-Based Perception of Fatal Crash Risks. *Accident Analysis & Prevention*, Vol. 151, 2021, p. 105962.
9. Karim, M. M., Y. Li, R. Qin, and Z. Yin. A Dynamic Spatial-Temporal Attention Network for Early Anticipation of Traffic Accidents. *arXiv Preprint arXiv:2106.10197*, 2021.
10. Ayoub, J., X. J. Yang, and F. Zhou. Modeling Dispositional and Initial Learned Trust in Automated Vehicles With Predictability and Explainability. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 77, 2021, pp. 102–116.
11. Shariff, A., J.-F. Bonnefon, and I. Rahwan. Psychological Roadblocks to the Adoption of Self-Driving Vehicles. *Nature Human Behaviour*, Vol. 1, No. 10, 2017, pp. 694–696.
12. Olaverri-Monreal, C. Promoting Trust in Self-Driving Vehicles. *Nature Electronics*, Vol. 3, No. 6, 2020, pp. 292–294.
13. Raats, K., V. Fors, and S. Pink. Trusting Autonomous Vehicles: An Interdisciplinary Approach. *Transportation Research Interdisciplinary Perspectives*, Vol. 7, 2020, p. 100201.
14. Choi, J. K., and Y. G. Ji. Investigating the Importance of Trust on Adopting an Autonomous Vehicle. *International Journal of Human-Computer Interaction*, Vol. 31, No. 10, 2015, pp. 692–702.
15. Shen, Y., S. Jiang, Y. Chen, E. Yang, X. Jin, Y. Fan, and K. D. Campbell. To Explain or Not to Explain: A Study on the Necessity of Explanations for Autonomous Vehicles. *arXiv Preprint arXiv:2006.11684*, 2020.
16. Ha, T., S. Kim, D. Seo, and S. Lee. Effects of Explanation Types and Perceived Risk on Trust in Autonomous Vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 73, 2020, pp. 271–280.
17. Goodman, B., and S. Flaxman. European Union regulations on Algorithmic Decision-Making and a "Right to Explanation."*AI Magazine*, Vol. 38, No. 3, 2017, pp. 50–57.
18. Zablocki, É., H. Ben-Younes, P. Pérez, and M. Cord. Explainability of Vision-Based Autonomous Driving Systems: Review and Challenges. *arXiv Preprint arXiv:2101.05307*, 2021.
19. Simonyan, K., A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Proc., Workshop at International Conference on Learning Representations*, Citeseer, Banff, Canada, 2014.
20. Mahendran, A., and A. Vedaldi. Salient Deconvolutional Networks. *Proc., European Conference on Computer Vision*, Springer, 2016, pp. 120–135.
21. Montavon, G., S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. Explaining Nonlinear Classification Decisions With Deep Taylor Decomposition. *Pattern Recognition*, Vol. 65, 2017, pp. 211–222.
22. Montavon, G., A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. Layer-Wise Relevance Propagation: An Overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 2019, pp. 193–209.
23. Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *Proc., IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 2921–2929.
24. Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proc., IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 618–626.
25. Chattopadhay, A., A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM + +: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. *Proc., IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, IEEE, New York, 2018, pp. 839–847.
26. Fu, R., Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li. Axiom-Based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. *arXiv Preprint arXiv:2008.02312*, 2020.
27. Muhammad, M. B., and M. Yeasin. Eigen-CAM: Class Activation Map Using Principal Components. *Proc., 2020 International Joint Conference on Neural Networks (IJCNN)*, Padova, Italy, IEEE, New York, 2020, pp. 1–7.
28. Bojarski, M., A. Choromanska, K. Choromanski, B. Firner, L. J. Ackel, U. Muller, P. Yeres, and K. Zieba. Visualbackprop: Efficient Visualization of CNNs for Autonomous Driving. *Proc., 2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, IEEE, New York, 2018, pp. 4701–4708.
29. Kim, J., A. Rohrbach, T. Darrell, J. Canny, and Z. Akata. Textual Explanations for Self-Driving Vehicles. *Proc., European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 563–578.

30. Cultrera, L., L. Seidenari, F. Becattini, P. Pala, and A. Del Bimbo. Explaining Autonomous Driving by Learning End-to-End Visual Attention. *Proc., IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Nashville, TN, 2020, pp. 340–341.
31. Sauer, A., N. Savinov, and A. Geiger. Conditional Affordance Learning for Driving in Urban Environments. *Proc., Conference on Robot Learning*, Zürich, Switzerland, PMLR, 2018, pp. 237–252.
32. Gallitz, O., O. De Candido, M. Botsch, and W. Utschick. Interpretable Feature Generation Using Deep Neural Networks and its Application to Lane Change Detection. *Proc., 2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Auckland, New Zealand, IEEE, New York, NY, pp. 3405–3411.
33. Liu, Y.-C., Y.-A. Hsieh, M.-H. Chen, C.-H. H. Yang, J. Tegner, and Y.-C. J. Tsai. Interpretable Self-Attention Temporal Reasoning for Driving Behavior Understanding. *Proc., ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, IEEE, 2020, pp. 2338–2342.
34. Xu, Y., X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos. Explainable Object-Induced Action Decision for Autonomous Vehicles. *Proc., IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 9523–9532.
35. Ren, S., K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. *Advances in Neural Information Processing Systems*, Vol. 28, 2015, pp. 91–99.
36. Krizhevsky, A., I. Sutskever, and G. E. Hinton. Imagenet Classification With Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, Vol. 25, 2012, pp. 1097–1105.
37. Tobii Pro. Tobii Pro Fusion, 2021. https://www.tobiipro.com/product-listing/fusion/. Accessed July, 2021.
38. Bylinskii, Z., T. Judd, A. Oliva, A. Torralba, and F. Durand. What do Different Evaluation Metrics Tell us About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 3, 2018, pp. 740–757.
39. National Highway Traffic Safety Administration. *Traffic Safety Facts Annual Report Tables*. US Department of Transportation National Highway Traffic Safety Administration, n.d.https://cdan.dot.gov/tsftables/tsfar.htm#. Accessed November, 2021.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.