Attention-Based Interrelation Modeling for Explainable Automated Driving

Zhengming Zhang, *Graduate Student Member, IEEE*, Renran Tian, *Member, IEEE*, Rini Sherony, *Member, IEEE*, Joshua Domeyer, and Zhengming Ding, *Member, IEEE*

***Abstract*—Automated driving desires better performance on tasks like motion planning and interacting with pedestrians in mixed-traffic environments. Deep learning algorithms can achieve high performance in these tasks with remarkable visual scene understanding and generalization abilities. However, when common scene-parsing methods are used to train end-to-end models, limitations of explainability in such algorithms inhibit their implementations in fully automated driving. The main challenges include algorithm performance deficiencies and inconsistencies, insufficient AI transparency, degraded user trust, and undermining human-AI interactions. This research aids the decision-making performance and transparency of automated driving systems by providing multi-modal explanations, especially when interacting with pedestrians. The proposed algorithm combines global visual features and interrelation features by parsing scene images as self-constructed graphs and using an attention-based module to capture the interrelationship among the ego-vehicle and other traffic-related objects. The output modules make decisions while simultaneously generating semantic text explanations. The results show that the fusion of the features from global frames and interrelational graphs improves decision-making and explanation predictions compared to two state-of-the-art benchmark algorithms. The interrelation module also enhances algorithm transparency by disclosing the visual attention used for decision-making. The importance of interrelation features on the two prediction tasks is further revealed along with the underlying mechanism of multitask learning on the datasets with hierarchical labels. The proposed model improves driving decision-making during pedestrian interactions with intelligible reasoning cues for building an appropriate mental model of automated driving performance for human users.**

***Index Terms*—Interpretable AI, automated driving, scene understanding, multi-task learning.**

Manuscript received 2 August 2022; revised 30 October 2022; accepted 15 November 2022. Date of publication 16 December 2022; date of current version 20 March 2023. This work was supported in part by the Toyota Collaborative Safety Research Center, and in part by the National Science Foundation under Grant 2145565. *(Corresponding author: Renran Tian.)*

Zhengming Zhang is with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: zhan3988@purdue.edu).

Renran Tian is with the Department of Computer Information & Graphics Technology, Indiana University Purdue University Indianapolis (IUPUI), Indianapolis, IN 46202 USA (e-mail: rtian@iupui.edu).

Rini Sherony and Joshua Domeyer are with the Collaborative Safety Research Center, Toyota Motor North America, Ann Arbor, MI 75024 USA (e-mail: rini.sherony@toyota.com; joshua.domeyer@toyota.com).

Zhengming Ding is with the Department of Computer Science, Tulane University, New Orleans, LA 70118 USA (e-mail: zding1@tulane.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2022.3229682.

Digital Object Identifier 10.1109/TIV.2022.3229682

I. Introduction

THANKS to the rapid development of modern AI techniques, automated driving is becoming closer to reality. Deep learning algorithms have had breakthroughs in many automated driving tasks, such as perception [1], [2], motion planning [3], [4], [5], and obstacle detection [6], [7]. Different technical solutions have been proposed and implemented to either break down automated driving into sub-tasks such as lane detection, obstacle avoidance [8], [9], and pedestrian intent prediction [10], [11], [12], or develop holistic systems to design an end-to-end network that executes multiple tasks simultaneously [13], [14], [15].

Even though multi-modal sensors provide alternative or supplementary methods for vehicle perception [16], [17], [18], [19], they often come with higher expenses in both computation and fabrication. On the other hand, high-resolution cameras are relatively low-cost while capturing detailed visual information such as traffic signs and pedestrian facial expressions. Thus, the mainstream solution for automated driving heavily relies on optical sensors, creating deep-learning algorithms using images and videos as inputs. However, despite the superior performance of modern computer-vision algorithms in driving decision-making, the "black box" property of these algorithms decreases the transparency, interpretability, and explainability of the underlying mechanisms that contribute to outcomes. Such limitations may inhibit the penetration of automated driving by affecting user attitudes and intention to use. This includes topics such as trust between the users and the automated driving system [20], [21], [22], [23], [24], human-AI teaming efficiency [25], [26], and the capability to diagnose and improve the algorithms from system-centric information [27], [28].

However, few studies focus on the interrelationship across traffic-induced objects to support explainable decision-making for automated driving. To the best of our knowledge, only Yao et al. [29] devised an attentive relation network to explore the relational features of the surrounding objects and predict pedestrian behaviors. Their model reconstructed the scenario using masks from a pre-trained semantic segmentation model. However, the model only implemented segmentation features and did not include local visual features, which may result in the model not fully reflecting the interrelationships among scene objects.

In this study, we propose an attention-based deep learning algorithm to predict feasible actions of the ego vehicle and

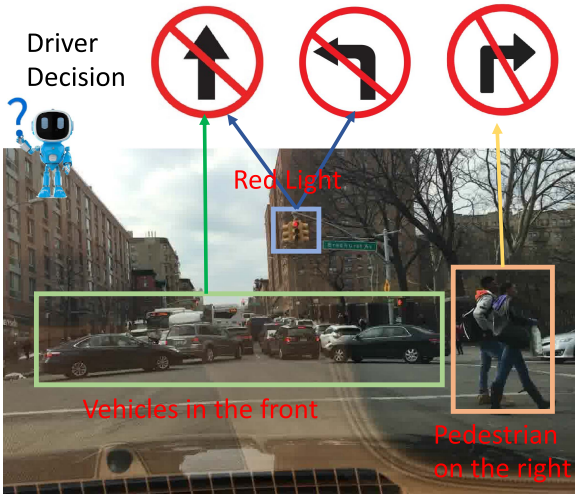

Fig. 1. A demonstration for our studied problem. Specifically, there are four potential driving actions (Left/right turn and forward/stop.) Given a driving scenario (dash-cam image), the model needs to infer the possible driving actions and explains the corresponding decision-making reasoning.

explain the corresponding decision-making reasoning, given a driving scene captured in an egocentric (dash-cam) image (in Fig. 1). The proposed method extends existing research by studying interrelationships among scene objects, including both visual and segmentation features, to support explainable automated driving. The direct use of the model is to deploy it in the level 2 and level 3 automated driving systems [30] to support driving decision-making in complicated urban settings, especially during pedestrian encounters. The outputs of the proposed module can mimic human driver decision-making in complex situations for both lateral and longitudinal controls and explain how the driving decisions are made automatically with visual and semantic feedback. The explanations can help to calibrate the trust in automated driving and improve user acceptance and performance facing such AI systems.

The key innovation of our work is the interrelation module, which encodes the interrelationship among the traffic-related objects into a fully connected graph. The pooled graph representation is fused with global visual features to create hidden embeddings for the traffic context. Finally, the embeddings are used for predicting driving actions with corresponding explanations using a non-linear classifier. The performance of our proposed model outperformed state-of-the-art models on two benchmark datasets. In addition, we demonstrate the underlying reasons for the performance improvement using a multitask learning mechanism. To sum up, our key contributions are highlighted below:

- We model the interrelationship among the traffic-related objects to extract discriminative context embeddings considering both visual and segmentation features.
- The proposed framework incorporates global and local contexts to predict driving actions and explanations of the decision-making process, where the performance surpasses the state of the art on two benchmark datasets.
- We show that multi-task predictions on action and explanation together boost the performance due to the hierarchical structure.

II. Related Works

Many studies have highlighted the importance of explainable automated driving models [3], [31], [32], [33], [34], [35], especially from the perspective of human-AI cooperation. On one hand, the unpredictable performance of automated driving substantially undermines user acceptance when explanations are not included. On the other hand, Koo et al. found that providing the underlying reasons for automated driving decisions was preferred by the drivers and led to a better AI-assisted driving performance [36]. One recent work found a significant effect of explainability on the perception, trust, and acceptance of general AI systems [37]. Thus, explainability in automated driving systems is a critical and promising research focus.

Since automated driving is a highly context-dependent task, the interpretation of AI algorithms involves describing the key contextual features and their effects on driving decisions. In one study, [38] proposed an object-centric architecture for traffic policy learning and found that it was superior when compared with object-agnostic methods on both simulated and real traffic. In another study, [31] devised a selector module identifying the action-induced objects and mapped the objects to explanations. The model proposed by [39] transforms the visual observation into natural language along with the driving actions. In addition to predicting driving actions with explanations, research [40], [41] has also predicted drivers' visual attention to assist with AI interpretability. The above-mentioned studies have shown that traffic-related objects may be a suitable medium to convey the context and convey the underlying causality of automated driving algorithms.

Although both one or many objects can explain the automated driving maneuvers to a great extent, modeling complex driving contexts by only relying on the existence and appearance of individual objects might overgeneralize the scenarios. Intuitively, the relationship between the key objects might make the context more discriminative and is critical for understanding traffic rules and social norms. In particular, the interrelationship is critical for understanding and following social and traffic rules. Despite some deterministic traffic rules like "stop at the right light" and "yield at the roundabout" that are well-defined and easier to be code, many other aspects of the object interrelationships, like the social norms, are more complicated and context-dependent. Those social rules are difficult to be manually engineered. We believe that a deep learning module specialized to learn the interrelationships is necessary to facilitate mining traffic and social rules.

III. The Proposed Algorithm

A. Problem Setup

Given a dash-cam image, our major objective is to predict feasible maneuvers of automated driving while explaining the

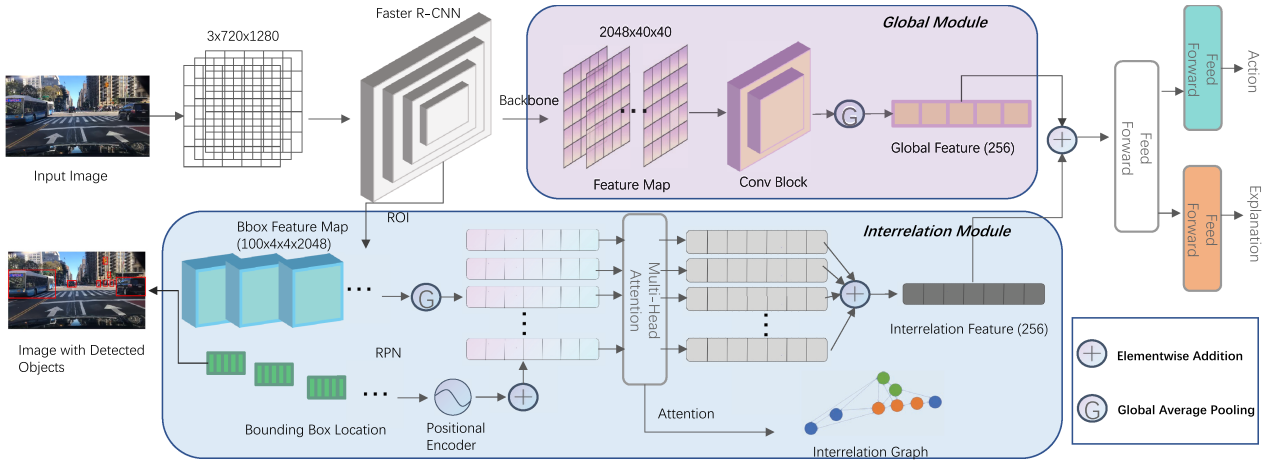

Fig. 2. Overview of the proposed model with the input as an RGB image, which is then fed into a Faster R-CNN for backbone features extraction and object detection. The main goal of the Faster R-CNN is to extract the global context features, which is its backbone feature, and the local object features, which are the average pool features after the ROI alignment for each detected bounding box. Eventually, two new proposed modules work together for driving decision-making and human explanation.

corresponding reasoning process, which is fundamental for motion planning that is interpretable. We define the problem as a multilabel classification problem: given a dashcam image I in some space χ, the goal is to determine feasible actions $A \in \Theta$ with explanations $E \in \{0,1\}^{n_e}$. Θ in our study has two types, depending on the evaluation datasets. Details of the datasets are provided in section IV-A.

1) In the BDD-OIA dataset [31], Θ is defined as $\{0,1\}^4$ (four binary independent action categories).
2) In the PSI dataset [42], Θ is defined as $\{0,1,2\}$ (one action category with three choices).

The n_e is the number of explanation categories. Mathematically, we aim to construct a mapping function:

$$\phi : \chi \rightarrow (A, E) \in \Theta \times \{0,1\}^{n_e}. \tag{1}$$

B. Framework Overview

Fig. 2 shows the overall architecture of our proposed model, which is composed of global and interrelation modules. Illustrated in the top part of the figure, the global module aims to capture the traffic context information at a higher level using the feature map of the entire input image. Our design of the global module extends work in [31], which is a simple but powerful way to capture the useful information of the backbone features.

Shown in the bottom half of Fig. 2, the interrelation module focuses on modeling the interrelationship among the detected scene objects. Details of this module will be provided in the next section. The embeddings from both modules are then fused into a final representation through elementwise addition. In the end, two classifiers with a shared fully connected layer are used for the multi-label classification.

We perform multitask learning, where the loss function is composed of two losses from both the predicted action and explanation. The multi-task loss function is formulated as:

$$L = L_{\text{action}} + L_{\text{explanation}}, \tag{2}$$

where L_{action} and $L_{\text{explanation}}$ are the summation of the binary cross entropy loss for each label.

C. Object Detetection Via Faster R-CNN

In order to detect the objects and extract the visual features of the images, we pre-trained a Faster R-CNN [43] with ResNet-50 structure [44] on the BDD-100 k dataset. The main goal of the Faster R-CNN is to extract the global context features, which are its backbone feature, and the local object features, which are the average pool features after the ROI alignment for each detected bounding box. Note we kept only 100 proposals with the largest confidence level from the RPN for simplicity.

D. Object Interrelation Discovery

Even though our proposed model does not explicitly use an existing graph neural network (GNN), the underlying concept is consistent with GNNs. We explain the intuition of our interrelationship model among traffic-related objects by comparing it to graph convolution networks (GCN) [45].

Generally, a graph is defined as a tuple of a node- and edge-set, where the node-set contains nodes in the graph and the edge-set describes the relationship among nodes. Graph-based modeling has been widely implemented. For instance, the relationships in a crowd are often modeled as social networks [46], where each node represents a person, and edges indicate certain relationships between people. Such a graph-based model can capture the interconnections among those in the group.

Similarly, driving scenes in dash-cam images can also be modeled with graphs. Extending our earlier work [12], we use traffic-related objects as nodes in the graph. The traffic-related objects in this study are the annotated objects in the BDD-100 K. There are 10 annotated classes in the dataset including pedestrian, rider, car, truck, bus, train, motorcycle, bicycle, traffic light, and a traffic sign. All the other objects are not considered traffic-related objects and are not included in the interrelation

modules. Driving a vehicle is a highly context-dependent task. The context includes both other road users (vehicles, bicyclists, or pedestrians) and road infrastructure (road position or traffic light), which are complicated and diversely perceived and processed by human drivers to infer upcoming actions. To better capture the scene features and traffic rules, the node features we designed could be decomposed into two parts:

1) **Bounding-box location features** include the pixel coordinates of the top-left and bottom-right corners and confidence level for the object detection, which are 5-dimensions vectors z_{li} for each node i.
2) **Visual features** z_{vi} are the corresponding global pooled features from the ROI (Region of Interests) for each object proposal i, extracted with a pre-trained backbone (Faster R-CNN in our implementation). Please refer to Section III-E for more details about proposals in the object detection phase.

Earlier work using graph models of the driving scenes usually defines edges manually based on certain assumptions [12], [47], like pixel distances among the objects. However, pre-determined edges are too simple to reflect relationship changes along with the context. Thus, we utilized a multi-head attention layer [48] to define the edge weights. Moreover, rather than the binary coding of the graph, percentage coding is adopted where any value between 0 to 1 is assigned as an edge weight where higher values indicate a greater ability to transmit a message.

Let $G = (V, E)$ be the interrelation graph, where each object detection proposal i is considered as a node $v_i \in V$. For all proposals, we designed a function $p : R^5 \rightarrow R^{2048}$ to map the location feature z_{li} to a vector h_{li} with length 2048. And then, the full node feature is the element-wise summation of the visual feature and transformed location feature, $z_{fi} = z_{li} + h_{li}$. The function p served as the positional encoder, commonly used in natural language processing (NLP) to incorporate features with positional information.

All nodes in the graph are fully connected, where the key k_i and query q_j determine the edge weights between node i and node j. In particular, $k_i = z_{fi}\theta_k$ and $q_j = z_{fj}\theta_q$. The node features are first transformed through a neural network Θ and then summed by the edge weights. Therefore, the hidden features of all nodes H_f^{t+1} at the multi-head attention layer $t + 1$ could be expressed as the following:

$$H_f^{t+1} = \text{softmax}\left(K^{t+1}Q^{t+1^\top}\right)H_f^t\Theta^{t+1}, \tag{3}$$

$K^{n\times n_k}$ and $Q^{n\times n_k}$ are the matrices of row-stacked keys k and queries q, n, and n_k are the numbers of nodes, and the dimension of keys and queries, and the softmax$(.)$ is a row-wise softmax operator. The node hidden feature H_f^0 at the first layer is Z_f. The above-mentioned processes are very similar to the graph convolution network, except that the edge weights are learned from the similarity measure of the features represented by the products of keys and queries.

Since the task is a multi-label classification, we used a summation to read out the sequence into a single vector, h_{gk} representing the graph k. The output vector is considered an interrelation feature that concentrates the information of the traffic-related objects and their context-based relationships into an epitome.

E. Proposals Filtering Enhancement

We chose the Faster R-CNN to detect traffic-related objects. Though using a pre-trained detector saves many workforces and promotes an end-to-end framework, it came with flawed performance. Therefore, to deal with potential object detection errors, we use the confidence levels of the object detection outputs to devise a data augmentation strategy. Among all the detection proposals from any pre-trained detectors, we draw a subset with the selection probability being proportional to the object detection confidence level. Since each proposal is selected independently, the distribution of the proposal i being selected follows Bernoulli(p_i), where p_i is the proposal i's confidence level.

To better control the selection of object proposals and balance the contributions between the global and interrelation modules, we manipulate the confidence level by calculating $(p_i)^\lambda$ to replace the original p_i, where $\lambda \in [0, \infty]$.

- As the λ value goes bigger towards infinity, the likelihood of being selected will reduce to zero, even for the high confident proposals. In this case, the global module will contribute more to the final prediction.
- In contrast, regardless of the confidence level, all proposals will be selected when the λ goes to zero, and the contributions from the interrelation module will increase.

By changing λ, the most appropriate object proposals can be selected. The proposed data augmentation guides the model to prevent excessively noisy detection by keeping high-confidence objects (true positives) and reducing low-confidence objects (false positives).

IV. Experiments

In this section, we elaborate on the two abovementioned datasets for action and explanation prediction. Then we discuss about the implementation details.

A. Datasets

BDD-OIA dataset [31] is a subset of BDD-100k [49], which is a large-scale, diverse driving video database. The video clips were recorded from a dashcam in front of the vehicle running in the naturalistic road environment. The original BDD-100 K dataset contains various driving scenes, including different weather and traffic conditions. While the BDD-100 k dataset is built for object detection, semantic segmentation, and other classic computer vision tasks, the BDD-OIA dataset focuses on the prediction of vehicle actions and driving-decision explanations.

Regarding the vehicle actions, the BDD-OIA dataset provides annotation of four available maneuvers (move forward, stop/slow, turn/merge left, and turn/merge right) for each image. In addition, the dataset also categorizes the possible explanations for the driving decisions into twenty-one classes and provides the corresponding explanations for each image. Both action and

TABLE I
STATISTICS OF BDD-OIA AND PSI DATASET

Dataset	Action	Number of Frames	#Reasoning
BDD-OIA	Forward	12,491	21 (human-defined)
	Stop/Slow Down	10,432	
	Turn Left	5,902	
	Turn Right	6,541	
PSI	Maintain Speed	5,800	29 (k-means clustered)
	Slow Down	4,925	
	Stop	1,177	

explanation predictions are multi-label classifications, i.e., each image could have more than one positive label for both actions and explanations. There are 22,924 images with annotations in the BDD-OIA datasets (Table I). Moreover, compared to other vehicle dashcam-based datasets (BDD-100 k, KITTI [50], and Cityscapes [51], there are much more pedestrians and vehicles appearing in the BDD-OIA, indicating the high complexity of the traffic context. As there are four action labels for each image, we first evaluate the action prediction performance using an F1 score for each specific action. Then, a sample-wise overall F1 score and a class-wise average F1 score are calculated to evaluate action and explanation prediction tasks.

PSI dataset [42] contains 110 15-second driving videos captured by a dashcam. The videos were randomly selected from a large naturalistic driving dataset [52] with potential conflicts between vehicles and pedestrians. The original data were collected from 116 human drivers continuously for one year. Twenty-four human drivers annotated each video for driving actions ("maintain speed," "slow down," and "stop") at keyframes based on their scene understanding. In addition to the driving decisions, human drivers describe the corresponding reasoning explanations. We used a pre-trained BERT [53] to map the explanation sentences into semantic embeddings and then used K-means to cluster the sentences into 29 explanation categories. In other words, each frame has one driving decision and its corresponding explanation categories. Note that the driving action in PSI is a single selection label which is different from the BDD-OIA. There is a total of 11,902 keyframes in the dataset, with the details shown in Table I.

All samples are split into training, validation, and test sets with a ratio of 7/1/2. Since the action annotation is a single-label classification, we first evaluate the action prediction performance using accuracy for each specific action. Then, we use overall prediction accuracy and class-wise average accuracy to evaluate the action prediction task, and sample-wise overall F1 score and the class-wise average F1 score for the explanation prediction task.

B. Implementation Details

First, the global module used the backbone feature map from a frozen Fast R-CNN with ResNet-50 to capture the entire scene. Next, we added two convolution blocks (convolution layer and ReLU activation functions) to lower the dimensionality and spatial size (256x7x7) and then feed it into an average pool layer to get the final representation for the global features.

The design of the interrelation module simplifies the scene context by focusing on the relationships among driving-related objects. The location and feature maps of the proposals are obtained from the ROI and RPN (Region Proposal Network) heads of the Faster R-CNN. In our implementation, the number of proposals is limited to 100 instead of the standard setting, which is 300. The feature maps of the proposals are then average pooled to vectorize the features. We add a fully connected layer as a positional encoder to transform the location feature into a vector with 2048 dimensions and then element-wisely add to the pooled visual features. The combined proposal feature thus contains both visual and position information. The multi-head attention block contains three multi-head attention layers. Each layer's value, key, and query features are transformed from the proposal features through a fully connected network with a ReLU activation function, respectively. The fully connected layer in the multi-head attention layer also reduces the dimensionality from 2048 to 256. The features generated from the multi-head attention block represent the interrelation graph.

After obtaining the features from the global and interrelation module, we used element-wise summation to fuse the features into one vector. Then, a fully connected layer with the ReLU activation function transformed the fused feature, before two separated fully-connected layers were applied as the classifier for action and explanation prediction.

The training procedure is different for single-module training and dual-module training. For single-module training, we trained the network with either the global module or the interrelation module from the random initialization. However, when training the entire network with dual modules, we initialized the weights of the global module with a pre-trained model and then trained the entire network. All models are trained for 50 epochs with no weight decay. The learning rate is set as 0.001 with the Adam solver.

V. RESULTS

In this section, we firstly show the comparison results of our proposed models with existing algorithms. Then, we discuss the influence of hyper-parameter in data augmentation, which could be considered an ablation study. At last, we demonstrated the underlying reason why such multitasks learning alleviated the performance by examining the intrinsic correlation between two tasks.

A. Comparison Results

We compare our proposed model with the state-of-the-art explainable model, OIA[1] to evaluate the performance. In addition, we add another baseline for comparison that is composed of a ResNet-101 pre-trained on ImageNet [54] to extract features and then followed by classifiers.

[1]Note that we cannot reproduce the performance of OIA as reported in the paper, so we include our reproduced results and the reported results.

TABLE II
ACTION AND EXPLANATION PREDICTION PERFORMANCE COMPARED AMONG OUR AND EXISTING MODELS ON BDD-OIA DATASET. IN THE TABLE, THE PERFORMANCE OF OIA* IS REFERENCED FROM THE PAPER, AND THE OIA IS REPRODUCED. DRIVING ACTION LEVELS DENOTE "MOVE FORWARD" (F), "STOP/SLOW DOWN" (S), "TURN/CHANGE LANE TO THE LEFT" (L), AND "TURN/CHANGE LANE TO THE RIGHT" (R). (THE BEST PERFORMANCE AMONG ALL MODELS ARE UNDERLINED, WHILE THE BEST PERFORMANCE AMONG ALL REALIZED MODELS ARE BOLD.)

Method	$F^{\dagger}$	$S^{\dagger}$	$L^{\dagger}$	$R^{\dagger}$	action $F1^{\star}_{all}$	action $mF1^{*}$	explanation $F1^{\star}_{all}$	explanation $mF1^{*}$
Baseline	0.755	0.607	0.098	0.108	0.601	0.392	0.331	0.18
OIA*	0.829	0.781	0.630	0.634	0.734	0.718	0.422	0.208
OIA	0.792	0.742	0.594	**0.627**	0.705	0.689	0.501	0.293
Ours	**0.802**	**0.753**	**0.619**	0.625	**0.722**	**0.701**	**0.537**	**0.335**

Note that † means F1 score for each action. * showsn sample-wise overall F1 score. * denotes class-wise average F1 score.

TABLE III
ACTION AND EXPLANATION PREDICTION PERFORMANCE COMPARED AMONG OUR AND EXISTING ALGORITHMS ON PSI DATASET

Method	MaintainSpeed†	SlowDown†	Stop†	action $Acc^{\star}_{all}$	action $mAcc^{\star\star}$	explanation $F1^{*}_{all}$	explanation $mF1^{**}$
Baseline	0.540	**0.774**	0.537	0.635	0.617	0.178	0.119
OIA	0.693	0.622	0.463	0.643	0.593	0.189	0.110
Ours	**0.713**	0.769	**0.615**	**0.712**	**0.699**	**0.278**	**0.191**

Note that † denotes the accuracy for each action. * means sample-wise overall accuracy. ** shows class-wise average accuracy. * is sample-wise overall F1. ** represents class-wise average F1.

On the BDD-OIA dataset, both our model and the reproduced OIA model outperform the baseline model significantly, as shown in Table II. Moreover, our proposed model outperforms the reproduced OIA model in both action and explanation predictions in most of the metrics, with about 2% and 10% improvements in terms of both F1 scores, respectively. The reproduced OIA model performs worse than the results shown in the original paper [31], but better for explanation prediction, which may be caused by different training strategies that we were not able to reproduce. In comparison with the original OIA model, the proposed model performs slightly worse in action prediction measured by both F1 scores (about 2% decrease), but has better explanation prediction capability (27% and 61% in two F1 scores respectively).

We also trained the baseline, OIA, and our proposed models using the PSI dataset to compare the performance, with the results shown in Table III. On this dataset, the proposed model outperforms both the baseline and the OIA algorithm significantly, with two main results:

- Driving action prediction accuracy has improved from around 60% with OIA to 70% with the proposed model, indicating promising results to support driving decision-making.
- Explanation prediction performance is worse using the PSI dataset for all models compared with the results on the BDD-OIA dataset. PSI data focuses on more complex driver-pedestrian interaction scenes with more complicated scene explanations. In this case, the proposed model still has 47% to 74% higher F1 scores compared with the OIA model trained on the same dataset.

B. Parameter Analysis

As mentioned above, the hyper-parameter λ in the data augmentation decides the number of selected proposals from the Faster R-CNN outputs and balances the contributions between the global and interrelation modules. When the λ goes to infinity, even the likelihood of being drawn from the highly confident proposals will shrink to zero, which leaves the global modules to do the work. In contrast, regardless of the confidence level, all proposals will be selected when the λ goes to zero. We test the proposed model when λ equals 0, 0.25, 1, 1.25, and infinity. To evaluate the performance of the interrelation module by itself, we add that into the experiments as well. The results of the models with different values of λ and with a single module are shown in Table IV.

As we see, when $\lambda = 1$, the balanced dual-module model has the best performance on all the action predictions. And selecting all proposals ($\lambda = 0$) results in the best performance on the explanation prediction. Several key findings include:

- The global module is more important in driving scene understanding in terms of both action and explanation prediction. The abundant information extracted from the whole image makes the global module perform better compared with the interrelation module.
- Compared with the global module itself, adding local object features and the interrelationships can improve action and explanation prediction performance. The dual-module models with different λ values almost always perform better than the global module itself.
- For action prediction, the balanced dual-module model performs the best, meaning that although certain levels of interrelation information are helpful, adding too many objects may distract the algorithm from the key features and thus reduce the action prediction capability.
- More interrelation information is important for continuous improvement of explanation prediction performance. This result may be caused by the fact that human drivers use more features than the trained learning algorithm for driving decision-making, and more interrelation information can help the algorithm to capture these human-used cues.

TABLE IV
ACTION AND EXPLANATION PREDICTION PERFORMANCE UNDER DIFFERENT λ IN THE DATA AUGMENTATION. DRIVING ACTION LABELS DENOTE "MOVE FORWARD" (F), "STOP/SLOW DOWN" (S), "TURN/CHANGE LANE TO THE LEFT" (L), AND "TURN/CHANGE LANE TO THE RIGHT" (R). (THE BEST PERFORMANCE AMONG ALL MODELS IS BOLD)

height λ or Module	$F^{\dagger}$	$S^{\dagger}$	$L^{\dagger}$	$R^{\dagger}$	action $F1^{\star}_{all}$	action $mF1^{*}$	explanation $F1^{\star}_{all}$	explanation $mF1^{*}$
0 (select all)	0.799	0.741	0.600	0.618	0.714	0.690	**0.558**	**0.356**
0.25	0.798	0.744	0.612	0.620	0.717	0.693	0.541	0.344
1	**0.802**	**0.753**	**0.619**	**0.625**	**0.722**	**0.701**	0.537	0.335
1.25	0.801	0.744	0.602	0.619	0.716	0.691	0.549	0.332
∞ (Global)	0.801	0.733	0.573	0.581	0.702	0.673	0.426	0.228
Interrelation	0.782	0.722	0.476	0.557	0.670	0.634	0.342	0.178

Note that † means F1 score for each action. * is the sample-wise overall F1 score. * denotes the class-wise average F1 score.

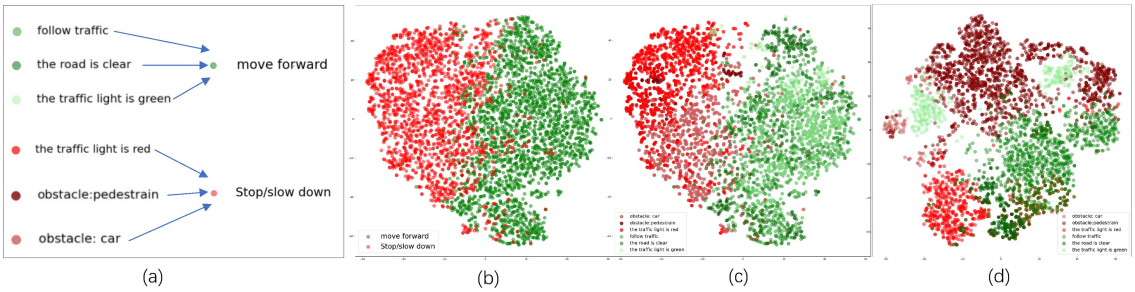

Fig. 3. (a) shows the hierarchical relationship between the selected explanation and action categories. (b), (c) and (d) are the t-SNE visualizations of the hidden features, where (b) refers to the samples with positive action predictions in "move forward" and "stop/slow down," (c) and (d) refers to the samples with positive predictions in the selected explanations. (b) and (c) are generated using the balanced dual-module model with multi-task learning, and (d) is from the single-module model focusing on explanation prediction only.

To conclude, this experiment shows that interrelation features can supplement global features for better driving scene understanding. Although more interrelation features can always help the algorithm to learn human explanations, many of them are not useful for predicting driving actions. Such discrepancies reflect the fundamental differences between human and AI decision-making. Because the balanced dual-module model performs the best in action prediction, we set the λ as 1 in later analysis.

All models with two modules perform similarly because the data augmentation does not change the input of the global module. And the global module is loaded with pre-trained weights. Therefore, no matter how many proposals are fed into the interrelation module, the network could always rely on the global module. The global module outperformed the interrelation module (select all proposals) because it has much richer input information.

C. Multi-Task Learning Boosting

As shown in our results (Table IV), multi-task [55] learning improves performance on both action and explanation predictions. Under the intuition that explanation of a certain scene shall help to infer the corresponding action, we believe that multi-task learning on a dataset with hierarchical labels, like BDD-OIA and PSI, enables the model to capture the hierarchical relationship between the explanation and action prediction tasks, and can supplement extra information when inferring one based on the other.

To further investigate how the hierarchical structures of the dataset and model cause the performance-boosting phenomenon, one critical piece of evidence needed is that the model captures the hierarchical relationship between tasks. We extracted the last shared hidden feature from our balanced dual-module model ($\lambda = 1$, trained on BDD-OIA), i.e., the second to the last fully connected layers (Fig. 2). Then, we used t-SNE [56] to transform the dimension from 64 to 2 for visualization. For comparison, we also trained the single-module model for explanation prediction.

To simplify the comparison, we chose six explanation categories corresponding to two action categories for demonstrations. These explanations and actions and their relationship are shown in Fig. 3(a). The main hypothesis is that for samples with position predictions of certain driving actions, "Move Forward" or "Stop/Slow Down," they shall also predict the corresponding explanations, so that we show that the algorithm learns the two tasks simultaneously.

Fig. 3(b) shows the distribution of samples with positive predictions of the two driving actions, and Fig. 3(c) shows the distribution of samples with positive predictions of the six explanations. The results show consistent patterns between the two results that samples for each action occupy one side of the hear-shape manifold ("Move Forward" to the right and

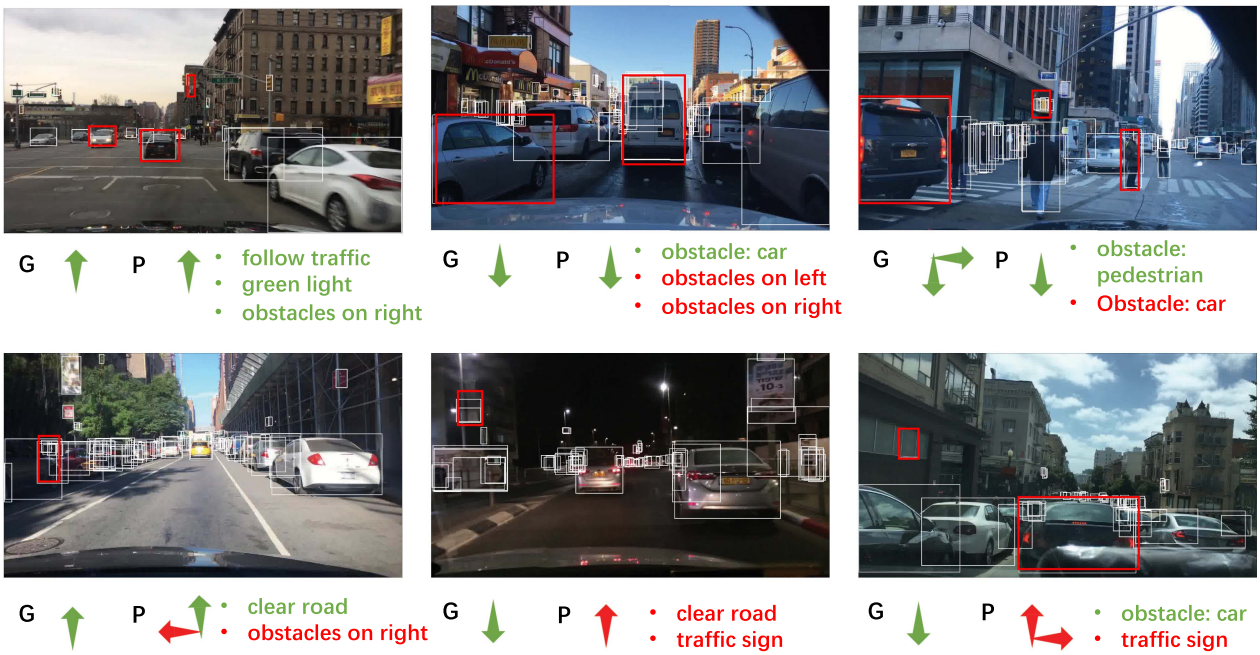

Fig. 4. Selected examples of action and explanation prediction on BDD-OIA dataset. G denotes the ground-truth annotation, and P shows our predicted result. Green and red arrows represent correct and wrong predictions, respectively. The green explanation predictions are True Positives, and the red ones are False Positives. Red bounding boxes are induced objects reflected by the model, while the white bounding boxes are the proposals from Faster R-CNN.

"Stop/Slow Down" to the left), with predictions of corresponding explanations as well. Thus, the hidden features in multi-task learning incorporate the hierarchical relationship in the data, and a model may leverage such information to make better inferences.

For comparison, we also show the hidden features of the single-module model only focusing on predicting the six explanations in Fig. 3(a). Fig. 3(d) shows the distributions of sample embeddings with corresponding positive explanation predictions. The pattern looks very different from Fig. 3(c). In particular, the greenish dots (samples with explanations that should correspond to the "Move Forward" action) and the reddish dots (samples with explanations that should be corresponding to the "Stop/Slow Down" action) are mixed in Fig. 3(d), indicating that the samples are not predicting corresponding actions consistently (i.e., no straight line could separate the samples into two actions).

D. Case Study for Visual Attention Illustration

Since one main research goal is to improve the explainability of driving decision-making algorithms, the proposed model not only supports explicit explanations as final outputs but can also feed back the visual attention for generating these action and explanation prediction outputs. We visualize the attention from the first-layer multi-head attention for a better understanding of the interrelation module. Due to many proposals, it isn't easy to illustrate the interrelation graph. Instead, we draw the bounding boxes on the object proposals with high attention scores from the others by calculating $\{b_i | \#\text{ of attention}_{ij}^{>0.3} >= 5, \forall i \in B\}$, where b_i is the bounding box, $\text{attention}_{ij}^{>0.3}$ is the indicator function for the attention score between proposal i and j is greater than 0.3, and B is the set of bounding boxes. In other words, we are selecting objects that have relatively higher attention than at least five other objects, which are considered as highly influential ones.

The images in the first rows of Fig. 4 show cases when the model performs well. The results show that when the model valued the correct objects, such as the leading vehicles, obstacle vehicles or pedestrians, and traffic lights, the predictions of driving actions and explanations are more accurate. On the contrary, the second rows in Fig. 4 show some worse cases. The results clearly demonstrate that when visual attention is distracted by noisy and irrelevant objects, the predictions of driving actions and explanations are less accurate.

The capability of localizing algorithm visual attention can help us to understand the model performs better. Some observations show that the proposed model still suffers from noisy detection and is sometimes distracted by objects in the top left, which might be caused by the positional encoder. More importantly, the visualization of visual attention can generate direct feedback to common users about the algorithm decision-making process so that the algorithm transparency is improved.

VI. Conclusion

The explainability of automated driving decision-making algorithms may aid in their development and acceptance and should be improved along with the overall action prediction performance. Contextual scene features that are better modeled

and connected can help with this goal. This paper proposes an attention-based module to capture the interrelationship among traffic-related objects and then combines the interrelation module with a global module to build a dual-module multi-task algorithm that can predict driving actions and explanations simultaneously for given driving scene images. The interrelation module provides the possibility to enable the illustration of visual attention toward prediction outputs to further improve algorithm transparency.

We exhaustively tested the performance of the proposed model on two benchmark datasets, and our model outperformed the baseline and state-of-the-art models in predicting both ego-vehicle actions and their corresponding explanations. Additional experiments reveal the importance of interrelation features in predicting actions and explanations, and the underlying mechanisms for improving performance using multitask learning on datasets with hierarchical labels. The proposed model sheds light on developing automated driving algorithms with improved decision-making performance and explainability when faced with complicated driving scenes.

REFERENCES

[1] H. Zhu, K.-V. Yuen, L. Mihaylova, and H. Leung, "Overview of environment perception for intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 10, pp. 2584–2601, Oct. 2017.

[2] Z. Wang, J. Zhan, C. Duan, X. Guan, P. Lu, and K. Yang, "A review of vehicle detection techniques for intelligent vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2021.3128968.

[3] A. O. Ly and M. Akhloufi, "Learning to drive by imitation: An overview of deep behavior cloning methods," *IEEE Trans. Intell. Veh.*, vol. 6, no. 2, pp. 195–209, Jun. 2021.

[4] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 1343–1350.

[5] X. Tang et al., "Prediction-uncertainty-aware decision-making for autonomous vehicles," *IEEE Trans. Intell. Veh.*, vol. 7, no. 4, pp. 849–862, Dec. 2022, doi: 10.1109/TIV.2022.3188662.

[6] V. D. Nguyen, H. V. Nguyen, D. T. Tran, S. J. Lee, and J. W. Jeon, "Learning framework for robust obstacle detection, recognition, and tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1633–1646, Jun. 2017.

[7] S. Ramos, S. Gehrig, P. Pinggera, U. Franke, and C. Rother, "Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling," in *Proc. IEEE Intell. Veh. Symp.* (IV), 2017, pp. 1025–1032.

[8] L. Xie, S. Wang, A. Markham, and N. Trigoni, "Towards monocular vision based obstacle avoidance through deep reinforcement learning," Robot.: Sci. Syst. Workshop: Boston, MA, US, 2017.

[9] Z. Zhang et al., "Implementation and performance evaluation of in-vehicle highway back-of-queue alerting system using the driving simulator," in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 1753–1759.

[10] K. Saleh, M. Hossny, and S. Nahavandi, "Intent prediction of pedestrians via motion trajectories using stacked recurrent neural networks," *IEEE Trans. Intell. Veh.*, vol. 3, no. 4, pp. 414–424, Dec. 2018.

[11] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6262–6271.

[12] T. Chen, R. Tian, and Z. Ding, "Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 3103–3109.

[13] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2174–2182.

[14] S. Hecker, D. Dai, and L. V. Gool, "End-to-end learning of driving models with surround-view cameras and route planners," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 435–453.

[15] Y. Pan et al., "Agile autonomous driving using end-to-end deep imitation learning," in *Proc. Robot.: Sci. Syst.*, 2018, doi: 10.15607/RSS.2018.XIV.056.

[16] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, 2018, Art. no. 3337.

[17] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 537–547, Jan. 2022.

[18] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7077–7087.

[19] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11784–11793.

[20] Q. Zhang, X. J. Yang, and L. P. Robert, "Expectations and trust in automated vehicles," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–9.

[21] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[22] S. Sheng et al., "A case study of trust on autonomous driving," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2019, pp. 4368–4373.

[23] Z. Zhang, V. G. Duffy, and R. Tian, "Trust and automation: A systematic review and bibliometric analysis," in *Proc. Int. Conf. Hum.- Comput. Interaction*, 2021, pp. 451–464.

[24] Z. Zhang, R. Tian, and V. G. Duffy, "Trust in automated vehicle: A meta-analysis.," Cham, Switzerland, 2023, pp. 221–234. [Online]. Available: https://doi.org/10.1007/978-3-031-10784-9_13

[25] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–autonomy teaming: A review and analysis of the empirical literature," *Hum. Factors: J. Hum. Factors Ergonom. Soc.,* vol. 64, no. 5, 2020, Art. no. 0018720820960865.

[26] C. Stephanidis et al., "Seven HCI grand challenges," *Int. J. Hum.–Comput. Interaction*, vol. 35, no. 14, pp. 1229–1269, 2019.

[27] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proc. 40th Int. Conf. Softw. Eng.*, 2018, pp. 303–314.

[28] Z. Zhang, R. Tian, V. G. Duffy, and L. Li, "The comfort of the soft-safety driver alerts: Measurements and evaluation," *Int. J. Hum.–Comput. Interaction*, , pp. 1–11, 2022. [Online]. Available: https://doi.org/10.1080/10447318.2022.2146324

[29] Y. Yao, E. Atkins, M. J. Roberson, R. Vasudevan, and X. Du, "Coupling intent and action for pedestrian crossing behavior prediction," in *IJCAI*, pp. 1238–1244, 2021.

[30] M. Blanco et al., "Human factors evaluation of level 2 and level 3 automated driving concepts," NHTSA, Washington, DC, USA, Tech. Rep. DOT HS 812 182, 2015.

[31] Y. Xu et al., "Explainable object-induced action decision for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9523–9532.

[32] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2942–2950.

[33] D. Shin, "Why does explainability matter in news analytic systems? Proposing explainable analytic journalism," *Journalism Stud.*, vol. 22, no. 8, pp. 1047–1065, 2021.

[34] T. Jing et al., "Inaction: Interpretable action decision making for autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 370–387.

[35] D. Shin, "User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability," *J. Broadcast. Electron. Media*, vol. 64, no. 4, pp. 541–565, 2020.

[36] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, "Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance," *Int. J. Interactive Des. Manuf.,* vol. 9, no. 4, pp. 269–275, 2015.

[37] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *Int. J. Hum.- Comput. Stud.*, vol. 146, 2021, Art. no. 102551.

[38] D. Wang, C. Devin, Q.-Z. Cai, F. Yu, and T. Darrell, "Deep object-centric policies for autonomous driving," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 8853–8859.

[39] J. Kim, S. Moon, A. Rohrbach, T. Darrell, and J. Canny, "Advisable learning for self-driving vehicles by internalizing observation-to-action rules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9661–9670.

[40] S. Baee, E. Pakdamanian, I. Kim, L. Feng, V. Ordonez, and L. Barnes, "MEDIRL: Predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13178–13188.
[41] D. Gopinath et al., "MAAD: A model and dataset for "attended awareness" in driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 3426–3436.
[42] T. Chen et al., "PSI: A pedestrian behavior dataset for socially intelligent autonomous car," 2021, *arXiv:2112.02604*.
[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 91–99.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[45] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–14.
[46] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proc. Nat. Acad. Sci.*, vol. 99, no. suppl_1, pp. 2566–2572, 2002.
[47] B. Liu et al., "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 3485–3492, Apr. 2020.
[48] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[49] F. Yu et al., "BDD100K: A diverse driving video database with scalable annotation tooling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2, no. 5, 2018, pp. 2636–2645.
[50] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[51] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 3213–3223.
[52] R. Tian, L. Li, K. Yang, S. Chien, Y. Chen, and R. Sherony, "Estimation of the vehicle-pedestrian encounter/conflict risk on the road based on TASI 110-car naturalistic driving data collection," in *Proc. IEEE Intell. Veh. Symp.*, 2014, pp. 623–629.
[53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. 1*, 2019.
[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[55] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, 2018.
[56] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE.," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008, pp. 2579–2605.

Zhengming Zhang (Graduate Student Member, IEEE) received the M.S. degree in statistics from the University of California, San Diego, San Diego, CA, USA. He is currently working toward the Ph.D. degree in industrial engineering with Purdue University, West Lafayette, IN, USA. His research interests include human-computer interaction, human factors, and deep learning for intelligent transportation systems.

Renran Tian (Member, IEEE) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 2002 and 2005, respectively, and the Ph.D. degree in human factors from Purdue University,West Lafayette, IN, USA, in 2013. He is currently an Assistant Professor of computer information & graphics technology with the Indiana University Purdue University Indianapolis, Indianapolis, IN, USA. His research interests include human-centered computing, human-AI teaming, artificial intelligence, cognitive ergonomics, and autonomous driving. He was the recipient of the NSF CAREER Award in 2022 from the Human-Centered Computing program.

Rini Sherony (Member, IEEE) received the master's degree in electrical engineering. She is currently a Senior Principal Engineer with Toyota's Collaborative Safety Research Center, part of Toyota Motor North America in Ann Arbor, Michigan. She has extensive experience in active safety and automated driving research, system design, evaluation/planning and data analysis. At CSRC, she leads active safety, automated driving collaboration research and data analysis. Her responsibilities, include development of standardized test procedures, test targets, testing, sensor requirements, and benefit estimation, for active safety systems Road Departure Mitigation system. She has led the development of SAE's pedestrian/bicyclist test target standards and also is involved in SAE/ISO Automated Driving activities. In addition to SAE, she is a Member of the Institute of Electronics and Electrical Engineers and the Association for the Advancement of Automotive Medicine. She has been an Organizer for SAE ADAS to Automated Driving conference, SAE Govt. Industry, and SAE World Congress ADAS/Ad sessions. She has authored or coauthored more than 120 papers/publications and has been granted 18 U.S. patents. She was the recipient of SAE's 2019 Forest R. McFarland Award. She is a Board Member for University of Michigan's CCAT (Center for Connected and Automated Transportation), Center for Automotive Research's, and Association for the Advancement of Automotive Medicine's Technical Advisory Board.

Joshua Domeyer received the B.S. and M.S. degrees in psychology from Central Michigan University, Mount Pleasant, MI, USA, in 2011, and the Ph.D. degree in industrial and systems engineering from the University of Wisconsin-Madison, Madison, WI, USA, in 2021. He is currently a Principal Researcher with Toyota's Collaborative Safety Research Center, part of Toyota Motor North America, Ann Arbor, MI, USA. His research include include human factors and includes topics such as vehicle-pedestrian interaction, driver attention modeling, and trust in automation. He is currently the Chair of SAE's Safety and Human Factors Steering Committee, a U.S. Expert of the TC22/SC39/WG8 Human-Vehicle Interaction Standards Group, and a Member of the Transportation Research Board's Standing Committee on Human Factors of Vehicles.

Zhengming Ding (Member, IEEE) received the B.Eng. degree in information security, and the M.Eng. degree in computer software and theory from the University of Electronic Science and Technology of China, Chengdu, China, in 2010 and 2013, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA, in 2018. He is currently a Faculty Member Affiliated with the Department of Computer Science, Tulane University, New Orleans, LA, USA. His research interests include transfer learning, multi-view learning, and deep learning. He received the National Institute of Justice Fellowship during 2016–2018. He is currently an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.