

Explainable Object-induced Action Decision for Autonomous Vehicles

Yiran Xu Xiaoyin Yang Lihang Gong Hsuan-Chu Lin
Tz-Ying Wu Yunsheng Li Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego

{y5xu, x4yang, lgong, lhsuanch, tzw001, yul554, nuno}@ucsd.edu

Abstract

A new paradigm is proposed for autonomous driving. The new paradigm lies between the end-to-end and pipelined approaches, and is inspired by how humans solve the problem. While it relies on scene understanding, the latter only considers objects that could originate hazard. These are denoted as action inducing, since changes in their state should trigger vehicle actions. They also define a set of explanations for these actions, which should be produced jointly with the latter. An extension of the BDD100K dataset, annotated for a set of 4 actions and 21 explanations, is proposed. A new multi-task formulation of the problem, which optimizes the accuracy of both action commands and explanations, is then introduced. A CNN architecture is finally proposed to solve this problem, by combining reasoning about action inducing objects and global scene context. Experimental results show that the requirement of explanations improves the recognition of action-inducing objects, which in turn leads to better action predictions.

1. Introduction

Deep learning has enabled enormous progress in autonomous driving. Two major approaches have emerged. *End-to-end* systems [1, 2, 3, 4, 5] map the visual input directly into a driving action, such as “slow-down” or “turn”. *Pipelined* systems first detect objects and obstacles, and then use that information to decide on driving actions. Both approaches have advantages and shortcomings. End-to-end systems are theoretically optimal, since all visual information is used for decision-making. By the data processing theorem [6], intermediate decisions can only lead to loss of information and compromise end-to-end optimality. However, end-to-end predictors are complex, requiring large datasets and networks. Pipelines have the advantage of modularity, decomposing the problem into a collection of much smaller sub-problems, such as object detection,

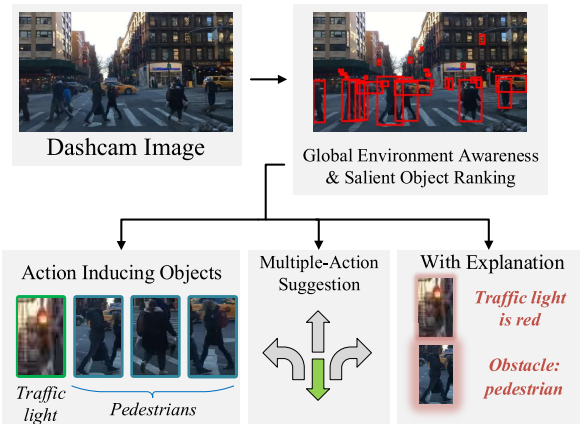


Figure 1. Top: while autonomous vehicles face complex scenes, composed of many objects, only a few of these are action-inducing. Bottom: each action-inducing object has an associated explanation for the related action. The arrows represent actions “move forward”, “turn left”, “stop/slow down”, and “turn right” (count-clockwise order). **Green** identifies the acceptable action.

trajectory analysis and planning, etc. This approach has spanned sub-literatures in 3D object detection [7, 8, 9, 10], segmentation [11], depth estimation [12, 13], 3D reconstruction [14], among other topics. Nevertheless, the failure of a single module can compromise the performance of the whole system [15].

Beyond driving performance, interpretability is a major concern for autonomous driving. End-to-end systems are notoriously poor in this regard given their black box nature. While they can be complemented with explanations [16, 17], these are not yet fully reliable. Pipelined systems are more amenable to forensic analysis, by analysing the performance of each module and assessing its potential contribution to a system failure. However, current pipelined approaches are not designed for this. Instead, each module becomes a computer vision problem of its own, sometimes with its own datasets and leaderboards. This makes it easy to loose track of the overall goal when evaluating performance. For example, further progress on the detection of

occluded pedestrians on sidewalks is unlikely to improve autonomous driving performance.

In this work, we advocate for the design of systems in between these two ends of the spectrum. This is inspired by how humans solve the problem. When deciding between slowing down or turning, humans do not employ a strict end-to-end strategy. Instead, they perform a certain amount of understanding and reasoning about scene objects. This, however, is far from full blown scene understanding. Instead, they only pay attention to objects directly related to the driving task. For example, when driving, most people pay little attention to pedestrians on the sidewalk, or parked cars, focusing instead on objects that could create hazard [18, 19]. We denote these as *action-inducing objects* because changes in their state, *e.g.* a pedestrian that crosses the street, trigger vehicle actions, *e.g.* “slow down” or “stop”. This is illustrated in Figure 1. While the scene contains many objects, only a few are action-inducing.

One advantage of focusing on action-inducing objects is that they define the set of explanations for driving actions. For example, in Figure 1, a car approaching an intersection slows down due to two action-inducing objects: a red traffic light and pedestrians crossing the street. More generally, *every action-inducing object has an associated explanation for the action it induces*. This implies that only a finite universe of explanations is required, and that explanations can be viewed as an auxiliary set of semantic classes, to be predicted simultaneously with the actions. This leads naturally to a multi-task problem, where the goal is to jointly predict actions and explanations, as shown in Figure 1.

In this work, we propose an architecture for this joint prediction. We start by introducing the BDD object induced action (BDD-OIA) dataset, for research in this problem. A limitation of current driving datasets [20, 21, 22, 23] is that they are too simple, *i.e.* it is frequently easy to determine the action to take. To address this problem, BDD-OIA only includes complicated scenes where multiple actions are possible. These are manually annotated for actions and associated explanations. BDD-OIA is complemented by a new architecture for joint action/explanation prediction, implemented by a multi-task CNN that leverages the Faster R-CNN to detect objects and a global scene context module to determine which of these are action-inducing, outputting the associated pair of actions and explanations.

The multi-task formulation has several nice properties. First, rather than an extra burden on classifier design, explanations become a secondary source of supervision. By forcing the classifier to predict the action “slow down” because “the traffic light is red,” the multi-task setting exposes the classifier to the causality between the two. This is much richer supervision than an image simply labelled as “slow down”. Second, unlike prediction heatmaps computed a posteriori [1, 3, 18, 19, 24, 25, 26, 27, 28], or the synthesis

of full blown natural language [3] with recurrent networks or generative systems, a finite set of explanations can be mapped into a classification task. Hence, the system can be explicitly optimized for generation of natural language explanations while posing a relatively simple learning problem. In fact, our experiments show that, under the proposed multi-task formulation, the generation of explanations *improves* the action prediction performance of the network. To the best of our knowledge, this is the first showing that explainable AI can improve the performance of a autonomous driving system. The proposed network is also shown to achieve good performance on the task of predicting driving commands and explanations, achieving state of the art results on BDD-OIA.

Overall, the paper makes four main contributions.

- A large dataset annotated for both driving commands and explanations.
- A new multi-task formulation of the action prediction problem that optimizes for both the accuracy of action commands and explanations.
- A CNN architecture for the solution of this problem, combining reasoning about action-inducing objects and global scene context.
- An experimental evaluation showing that the generation of explanations improves the decision making for actions, and both benefit from a combination of object-centric and global scene reasoning.

2. Related work

End-to-end learning for autonomous driving. End-to-end driving was first proposed in 1989, with the ALVINN system [29]. [30] later demonstrated the strong ability of CNNs to produce steering wheel commands. These systems are strictly end-to-end, using no explicit reasoning about objects. More recently a number of approaches [1, 2, 3, 4, 5] to end-to-end learning for vehicle control have taken advantage of both context and object features. However, many of these systems are trained on driving simulators [4, 5]. Despite progress in domain adaptation algorithms, systems trained on synthetic data tend to underperform when deployed in the real world. Our proposed network is an end-to-end system fully trained on real images, leveraging object detection and contextual reasoning.

Global-local contextual representations. Contextual relationships between global and local scene features are important cues for the identification of the important parts of a scene. Contextual reasoning has a long history in computer vision [31, 32, 33, 34, 35, 36, 37]. For example, [34] shows that multi-scale pooling improves the object detection performance of the Faster R-CNN[38] and [35] relies on contextual information to recognize visual relationships. However, contextual learning has received limited attention

in the autonomous driving literature. [5] proposed a selector of the most critical objects in the scene, but neglected the importance of global features altogether. [1, 3] considered instead features extracted from the whole scene, ignoring objects. Our proposed architecture accounts for both objects and context, exploring their relationships and detailed supervision, in the form of explanations, to separate action-inducing objects from objects unimportant for driving.

Attention mechanisms. Attention mechanisms have been widely utilized in neural networks [39, 40]. Attention maps are also used to visualize the inner workings of these networks [24, 25, 26, 27, 28]. In the autonomous driving realm, [41] developed a richer notion of attention on pixels that collaboratively contribute to the prediction. Studies of human attention, using eye tracking systems, have also been performed in [18, 19] to determine which scene regions capture the gaze of a driver. [1, 3] introduced the concept of visual attention maps for end-to-end driving. Instead of pixel-level attention, [5] proposed an object-level attention model. This could be seen as a weaker form of the now proposed idea of using objects to define actions.

Explanations. Textual explanations are sometimes used for insight on the network understanding of images or scene [16, 17, 42, 39]. For example, [39, 16, 17] generate text to explain either attention maps or network predictions. In the autonomous driving setting, Kim *et al.* [3] integrate textual generation and an attention mechanism with end-to-end driving. An attention-based video-to-text model is used to generate human understandable explanations for each predicted action. The formulation now proposed, based on action-inducing objects, enables one-hot encoded explanations. This eliminates the ambiguity of textual explanations, and improves action prediction performance.

Datasets. Several autonomous driving datasets, contain both real images or video and information from multiple sensors, including radar, LiDAR, GPS or IMU information. KITTI [20] was one of the earliest to be annotated with object bounding boxes, semantic segmentation labels, depth, and 3D point clouds. BDD100K [43] contains 100K videos annotated with image level labels, object bounding boxes, drivable areas, lane markings, and full-frame instance segmentation. Apolloscape [44] has 140K images, RGB videos and corresponding dense 3D point clouds with focus on 3D labeling and semantic segmentation. nuScenes [21] contains 1000 scenes with sensor information produced by camera, LiDAR, and radar. While large and richly annotated, none of these datasets addresses the detection of action-inducing objects. The dataset now proposed is derived from BDD100K but contains substantial extra annotations to enable this objective.

3. Joint Action and Explanation Prediction

In this section, we introduce the problem of jointly predicting and explaining object induced actions.

3.1. Definitions

Smart driving systems perform two major classes of actions. Actions in the first class are independent of other objects in the environment. For example, a car navigating on a desert freeway can simply decide to slow down to optimize energy consumption. These actions do not require sophisticated perception and are not hazardous. Actions in the second class involve reasoning about other objects on the road or its surroundings, as illustrated in Figure 1. While we refer to them as object induced actions, the definition of object could be abstract. For example, a lane change may be possible due to an open lane “object.” For the purposes of this work, any object or event that can be detected or recognized by a vision system is considered an object.

One of the interesting properties of object induced actions is their strong causal structure. For example, in Figure 1, the pedestrians that cross the street force the car to slow down. While there can be multiple causes for the action, *e.g.* the traffic light is also red, the cardinality of this set is relatively small. This implies that the action has a small set of possible explanations. If the car ran the intersection, it must have either not detected the pedestrians or the traffic light. While “corner” cases can always exist, *e.g.* the car failed to detect a broken tree limb in the middle of the road, these can be incrementally added to the set of explanations. In any case, because the set of objects and explanations is relatively small, the joint prediction of actions and explanations can be mapped into a factorial classification problem.

In this work, we consider the set of 4 actions commonly predicted by end-to-end driving systems [2, 5], and listed in the left of Table 1. These are complemented by the 21 explanations listed in the right side of the table. Different from previous works, we consider the classification of actions to be multi-label classification, *i.e.* we can have more than one choice. Mathematically, given an image I or a video V in some space \mathcal{X} , the goal is to determine the best action $A \in \{0, 1\}^4$ to take and the explanation $E \in \{0, 1\}^{21}$ that best justifies it. This is implemented by the mapping

$$\phi : \mathcal{X} \mapsto (A, E) \in \{0, 1\}^4 \times \{0, 1\}^{21}. \quad (1)$$

For instance, if the possible actions are “Stop” and “Change to the left lane”, then $A = [0, 1, 1, 0]^T$. The structure of the action and explanation label vectors is defined in Table 1. In summary, joint action/explanation prediction is a combination of two multi-label classification problems.

3.2. BDD-OIA Dataset

In the real world, driving is composed of long periods with very little to do (car simply “moves forward”) and rel-

Action Category	Number	Explanations	Number
Move forward	12491	Traffic light is green	7805
		Follow traffic	3489
		Road is clear	4838
Stop/Slow down	10432	Traffic light	5381
		Traffic sign	1539
		Obstacle: car	233
		Obstacle: person	163
		Obstacle: rider	5255
Turn left	838	No lane on the left	150
		Obstacles on the left lane	666
		Solid line on the left	316
	5064	On the left-turn lane	154
		Traffic light allows	885
Turn right	1071	Front car turning left	365
		No lane on the right	4503
		Obstacles on the right lane	4514
	5470	Solid line on the right	3660
		On the right-turn lane	6081
		Traffic light allows	4022
		Front car turning right	2161

Table 1. Action and explanation categories in the BDD-OIA dataset. Because actions are objectet induced, explanations are based on objects. Changing lanes to left/right is merged with turn left/right to avoid distribution imbalance. For the turn left/right rows, the upper sub-row presents statistics of changing lane to the left/right and the lower sub-row those of turning left/right. For these actions, explanations address why the action is not possible.

atively short periods where the driver must decide between a set of object induced actions. When decisions have to be made, they are more difficult if environments are complex, *e.g.* with road obstacles, pedestrian crossings, etc. Yet, driving datasets contain a relatively low percentage of such scenes. This can be seen from Table 2, which summarizes the average densities of pedestrians and moving vehicles per image of several datasets. The fact that these numbers are low suggests that most of the driving scenarios are relatively simple. Previous research also only predicts the action chosen by the driver [2, 4, 5], creating the false impression that only that action was possible. All of this, makes existing datasets poorly suited to study object induced actions. Beyond this, because these datasets are not annotated with explanations for object induced actions, they cannot be used to learn how to generate such explanations.

To address these problems, we selected a subset of BDD100K [43] video clips containing at least 5 pedestrians or bicycle riders and more than 5 vehicles. To increase scene diversity, these videos were selected under various weather conditions and times of the day. This resulted in 22,924 5-second video clips, which were annotated on MTurk for the 4 actions and 21 explanations of Table 1. We refer to this dataset as the *BDD Object Induced Actions* (BDD-OIA) dataset. Figure 2 shows examples of typical scenes in BDD-OIA. These are all complex driving scenes, where multiple action choices are frequently possible. There are also many objects, *e.g.* cars parked on the

Dataset	# pedestrians	# vehicles
BDD100K [43]	1.2	9.7
KITTI [20]	0.8	4.1
Cityscapes [22]	7.0	11.8
BDD-OIA	8.0	11.8

Table 2. Densities of pedestrians and vehicles per image in popular driving datasets (statistics based on training set). On average, the scenes of the proposed BDD-OIA dataset are more complicated than those of previous datasets.

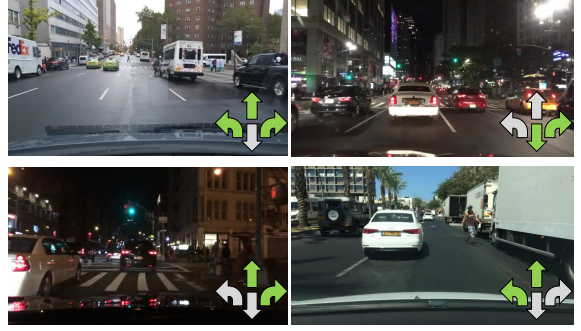


Figure 2. Scenes in BDD-OIA. The green arrows in the bottom right show the ground truth for possible actions.

side of the street, that are not action inducing and a few object inducing objects per scene. action-inducing objects can be other vehicles, pedestrians, traffic lights, or open lanes.

Table 1 summarizes the dataset content in terms of the 4 action classes and the 21 explanation categories. The coverage of actions is fairly balanced. In fact, our initial goal was to include the four BDD classes (“move forward,” “stop/slow,” “left,” and “right turn,”) plus the two novel classes of “change lane” to the left/right. However, there are very few opportunities to turn in BDD100K. To avoid a highly unbalanced dataset, we merged turns and lane changes. The coverage of the 21 explanation categories is a lot more unbalanced. The most probable is “Traffic light is green,” (7805 occurrences), while the rarest are “No lane on the left” (150) and “On the left-turn lane” (154).

4. Deep Learning Architecture

In this section, we propose a deep architecture for joint prediction and explanation of object induced actions.

4.1. Architecture overview

The prediction of object induced actions and their explanations requires a combination of several types of reasoning. In this work, we propose a deep network model based on several steps, which are illustrated in Figure 3. The network initially computes backbone features, which are fed to two modules. The local features $t_{l_i}, i = 1, 2, \dots, N$ are first produced by the RPN and ROI head layers of Faster R-CNN [38]. The Global module generates global features t_g by processing the size and dimension of the backbone fea-

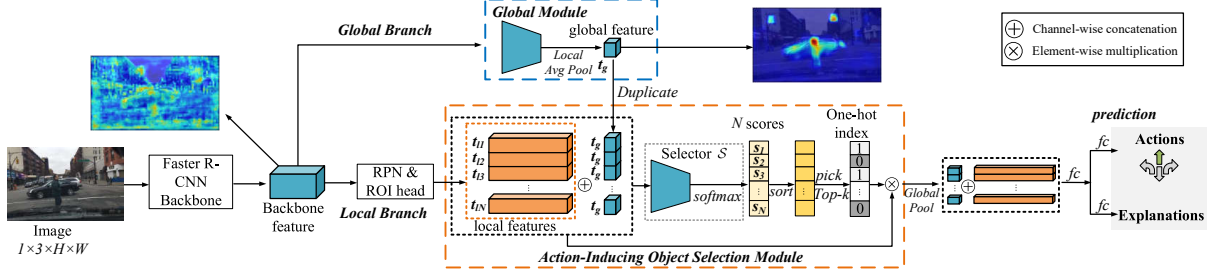


Figure 3. Architecture of the proposed network. The Faster R-CNN is used to extract backbone features, which are fed into a global and a local branch. The Global Module generates a global feature map that provides scene context, while the local branch captures the details of action-inducing objects. In the local branch, a selector module outputs a score for each object feature tensor and associated global context information. The top k action-inducing objects are selected and the features from the two branches are concatenated for action and explanation prediction. Two visualizations derived from the input image are also shown. The combination of local and global features and end-to-end supervision enables the network to reason about scene-object relationships and produce a global feature map more selective of action-inducing objects than the backbone feature maps.

tures so that they can be combined with the local features as well as modeling scene context and object-scene relationships. Then the local features and global features are fed to Action-Inducing Object Detection module to identify the action-inducing objects. These finally give rise to action \hat{A} and explanation \hat{E} predictions. The network is trained with a multi-task loss function

$$\mathcal{L} = \mathcal{L}_A + \lambda \mathcal{L}_E, \quad (2)$$

where $\mathcal{L}_A = \sum_{j=1}^4 L[\hat{A}_j, A_j]$ and $\mathcal{L}_E = \sum_{j=1}^{21} L[\hat{E}_j, E_j]$, A_j, E_j are the ground-truth labels for the j^{th} action and explanation, respectively, $L[.,.]$ is the binary cross entropy loss and λ an hyperparameter that controls the relative importance of action and explanation errors.

This formulation has several benefits. First, the multi-task training allows the explicit optimization of the network for the generation of explanations. This is likely to be more effective than deriving explanations by only a posterior, *e.g.* using only heatmaps that highlight image regions responsible for the prediction [27]. Second, because the generation of explanations is formulated as a classification problem, the optimization problem is fairly simple, requiring much less data than the training of natural language systems based on recurrent networks or generative language models [3]. Finally, due to the multi-task formulation, actions and explanations can benefit each other. By receiving explicit supervision for the fact that a car must slow down *because* a traffic light is red, the vision system faces a much simpler learning problem than one who is only told to slow down. It does not have to figure out on its own the causal relationship between the light being red and having to slow down.

4.2. Implementation Details

Global Module. This module generates global features t_g from the Faster R-CNN backbone features. It is composed of two convolutional layers with ReLU activation functions

plus a local average pooling operation. It reduces the dimensionality of the backbone features from 2048 to 256 and the spatial size of its features maps to 7×7 , to enable further joint processing of local and global features.

Action-Inducing Object Selection Module. This module is used to pick action-inducing objects from all object proposals produced by the Faster R-CNN. N local feature tensors t_{l_i} of size 7×7 are first extracted from the proposal locations and concatenated with the global feature tensor t_g to form an object-scene tensor $t_{(l+g)_i}$ per object. These tensors are then concatenated into a scene tensor of size $N \times c \times 7 \times 7$ where $c = 2048 + 256$. A selector \mathcal{S} then chooses the action-inducing objects from this tensor. \mathcal{S} is implemented with three convolutional layers and a softmax layer of N outputs, defining a probability distribution over the N objects. Probabilities are interpreted as action-inducing object scores. The k objects of largest score are then chosen as action-inducing and the associated object-scene tensors $t_{(l+g)_i}$ passed to the next network stage.

Predictions. These object-scene tensors are then globally pooled, and vectorized into a feature vector, which is fed to three fully connected layers to produce action predictions and explanations.

Object-scene relations. Together, the modules above allow the network to reason about scene-object relationships. The global module provides spatial context for where objects appear in the scene and global scene layout. It can be seen as an attention mechanism that combines the backbone feature maps to produce scene features informative of the location of action-inducing objects. This is illustrated in Figure 3, where we present an image, the average of the feature maps at the output of the backbone, and the average feature map after global module (the dimension reduced from 2048 to 256). While the backbone features have scattered intensity throughout the scene, the global feature maps are highly selective for action-inducing objects. This effect is complemented by the selector. Since the latter is learned

λ	F	S	L	R	action mF1	action F1 _{all}	explanation F1 _{all}
0	0.783	0.758	0.419	0.568	0.632	0.675	-
0.01	0.819	0.760	0.504	0.605	0.672	0.696	0.329
0.1	0.784	0.769	0.562	0.627	0.686	0.709	0.371
1.0	0.829	0.781	0.630	0.634	0.718	0.734	0.422
∞	-	-	-	-	-	-	0.418

Table 3. Action and explanation prediction performance as a function of the importance of each task (determined by λ) on the loss of (2). Labels denote “move forward” (F), “stop/slow down” (S), “turn/change lane to the left” (L), and “turn/change lane to the right” (R).

with supervision from the overall loss function \mathcal{L} , it chooses object-scene tensors that improve both the action prediction and explanation accuracy. This provides the global feature map with the supervisory signal needed to highlight the relevant objects. All the remaining proposals receive low score at the selector output and are discarded. This tremendously reduces the clutter created by objects that are unimportant to the action predictions.

5. Experiments

5.1. Setup

All experiments are based on the BDD-OIA dataset. Only the final frame of each video clip is used, leading to a training set of 16,082 images, a validation set of 2,270 and a test set of 4,572. The input size of images is $3 \times 720 \times 1280$. The Faster R-CNN is pre-trained on the annotated images from BDD100K [43] and frozen while the remainder of the network of Figure 3 is trained on BDD-OIA. The Adam optimizer is used with weight decay of 1×10^{-4} and initial learning rate α of 0.001. Training is performed for 50 epochs and α divided by 10 every 10 epochs. All experiments are evaluated with a standard metric, F1 score, with two variations considered.

$$F1_{all} = \frac{1}{|A|} \sum_{j=1}^{|A|} F1(\hat{A}_j, A_j), \quad (3)$$

averages the F1 score over all the predictions. Since the dataset is imbalanced, i.e. most of the actions are forward and slow, we further discuss the mean F1 score $mF1$ of each action j that only compute $F1(\hat{A}_j, A_j)$ for each sample.

The proposed network is compared to two other models: the ResNet-101 [45] (as Baseline) and the network of [5]. Because the latter were designed for tasks other than object induced action recognition, they are modified to support this task. ResNet-101 is pre-trained on ImageNet. In order to fit ResNet-101 architecture, the input size of images is resized to $3 \times 224 \times 224$. Its output layer is modified into 2 branches: a fully-connected (fc) layer that outputs 4 action categories, and a fc layer that outputs of 21 explanations. The network of [5] predicts driving actions. We add to it a new output branch for explanations. All models are trained with the loss of (2). The number of action-inducing objects is set to $k = 10$.

5.2. Interplay between Actions and Explanations

We started by investigating the impact of explanations on action prediction accuracy. For this, we varied the hyperparameter λ of (2), as summarized in Table 3. Note that $\lambda = 0$ corresponds to ignoring explanations during training and $\lambda = \infty$ to ignoring action predictions. Interestingly, the network trained uniquely to predict actions ($\lambda = 0$) has the weakest action prediction performance of all the models. Significant gains (an increase of action F1_{all} score from 0.675 to 0.734) are achieved when $\lambda = 1$, i.e. when explanations are given as much weight as actions. This model also has top performance for all action classes. This shows that explanations are not only useful, *but improve the performance of autonomous driving system*. We believe that this is the first showing that explainable AI systems can outperform uninterpretable systems in the vision literature.

Two properties of the proposed explanations justify this observation. First, the set of explanations is finite and defined based on objects. This, in turn, enables the robust learning of the explanation system from a limited set of examples. Open-ended explanation systems, based on natural language synthesized by recurrent models, lack this property. Second, and even more critical, the explanations of an object induced action recognition system are based on causal relations between objects, e.g. “stop because pedestrians are crossing the street”. This helps the system learn about object-scene relationships, e.g. figure out what to localize in the global feature map and relate local to global features, enabling a better identification of the action inducing objects and, consequently, simplifying action predictions. In the absence of explanations, the system has to figure out all these relationships by itself.

In summary, for the prediction of object induced actions, the addition of explanations is manageable and provides direct supervision about the causality of objects and induced actions that significantly simplify the learning problem. This can, in fact, be seen from the results of Table 3. Note that the addition of explanations produces a much larger gain for actions L and R, the classes of smaller representation in the dataset (see Table 1), than for actions F and S, the classes of larger representation (a ratio of 2:1 compared to L and R). This shows that, as the number of training examples declines and learning has more tendency to overfit, the regularization due to explanations produces

models	F	S	L	R	mF1	F1 _{all}	explanation mF1	explanation F1 _{all}
only local branch	0.760	0.649	0.413	0.473	0.574	0.605	0.139	0.351
only global branch	0.820	0.777	0.499	0.621	0.679	0.704	0.206	0.419
random selection in Selector	0.823	0.778	0.499	0.637	0.685	0.709	0.197	0.413
select top-5	0.821	0.768	0.617	0.625	0.708	0.720	0.212	0.416
select top-10	0.829	0.781	0.630	0.634	0.718	0.734	0.208	0.422

Table 4. Action and explanation prediction performance using global and local features. “Only local branch” (“Only global branch”) means that the network ignores global (local) features, “random Selector” that object features are chosen randomly, and “Select top- k ” that the selection module chooses the k objects of highest score.



Figure 4. Examples of network predictions, objects selected as action-inducing, and explanations. Yellow bounding boxes identify the objects detected by the Faster R-CNN, while red bounding boxes identify the objects selected as action-inducing by the proposed network. “G” stands for ground truth and “P” for prediction. For explanations, green indicates true positives, red false positives, and gray false negatives (*i.e.* valid explanations not predicted).

larger gains in action prediction performance.

5.3. Interplay between Local and Global Features

We next tested the importance of combining local and global reasoning. Table 4 summarizes a series of ablation experiments with different combinations of local and global features. We started by evaluating a model that only uses the local features derived from the Faster R-CNN detections. This achieved the worst performance of all models tested, for both actions and explanations. The action prediction is highly depended on spatial information without which the accuracy will drop a lot. We next considered a network using only global features and a network that picks the features from $k = 10$ random objects. While global features performed substantially better than local features, their performance was slightly weaker than that of the random selection. This suggests that it is too difficult to predict actions from all the Faster R-CNN object detections. The much im-

proved performance of global features supports the claim that they enable reasoning about the action-inducing scene parts. In fact, global features produced the best explanations of the three methods. Nevertheless, the slightly better action predictions of random object selection indicate that it is important to consider the objects in detail as well.

Given all this, it is unsurprising that the combination of the two feature types resulted in a significant additional performance gain, achieving the overall best results on the two tasks. This supports the hypothesis that action prediction requires reasoning about object-scene interactions. While both the selection of top 5 and top 10 objects, based on combination of local and global features, outperform all models of a single feature type, the number of objects has a non-trivial effect on network performance. In particular, better results were obtained with 10 than 5 objects. This confirms that BDD-OIA scenes are complex. On the other hand, the number of objects only had a marginal effect on explana-

models	F	S	L	R	mF1	F1 _{all}	explanation mF1	explanation F1 _{all}
Baseline	0.755	0.607	0.098	0.108	0.392	0.601	0.180	0.331
local selector [5]	0.810	0.762	0.600	0.624	0.699	0.711	0.196	0.406
ours	0.829	0.781	0.630	0.634	0.718	0.734	0.208	0.422

Table 5. Comparison of different models.

Outputs	F	S	L	R	mF1	F1 _{all}
single action	0.791	0.636	0.133	0.261	0.455	0.715
multiple action	0.795	0.680	0.522	0.594	0.648	0.665

Table 6. Single v.s. multiple action prediction. Single action prediction outputs a single action label given the input image. Multiple action outputs multiple labels.

tion performance. In fact, the model with only global features produced explanations of nearly equivalent quality to those of the entire network. This suggests that explanations mostly benefit from contextual reasoning.

5.4. Model comparisons

Table 5 compares the proposed network to the baseline and the method of [5]. The baseline is a purely global method, which predicts actions without extracting object features. It has the worst performance among all methods. This is further evidence for the importance of combining local and global features for action prediction. The model of [5] can be thought of a purely local selector, which uses no global features. Its performance is weaker than the proposed network, and similar to the random selection model of Table 4. Not surprising that this selector lacks the capacity for global reasoning. The gains of the proposed network show that object induced action recognition benefits from the analysis of contextual object-scene relationships.

5.5. Single vs. Multiple Action Predictions

Existing datasets assume that a single action prediction, that chosen by the driver, is possible at each instant. To investigate how this affects action prediction performance, we compared multiple and single action predictions. Single prediction ground truth is computed from IMU data in the original BDD100K dataset, which contains 11,236 training and 3,249 testing images. The network of Figure 3 is modified to produce a single action prediction, by addition of a softmax layer. This is compared to the original model, which can predict multiple actions. Table 6 shows that the performance of each action category is worse for single action predictions. This is for two reasons. First, the IMU labels exacerbate the class imbalance of the dataset. Among training images, there are 6,773 F, 4,258 S, 111 L, and 94 R labels. Seriously imbalanced data lead to models that predict F and S all the time. Second, single labels are deceiving. The fact that the driver chose F does not mean that it was not possible to chose L or R. In result, IMU labels are not ground truth for *possible actions*, they mostly reflect driver *intent*. The only conclusion possible from an F label is that the driver *wanted* to keep moving forward

and was not forced to stop, not that F was the only possible action. Again, because a driver typically chooses F and S much more frequently than L or R, the model is encouraged to always predict F or S. In summary, IMU labels encourage autonomous driving vehicles that do not know when it is possible to turn. The introduction of multiple action predictions substantially increases the number of examples per category, mitigating the data imbalance, and creating a lot more examples of scenes with turn labels, mitigating the turn-aversion problem.

5.6. Qualitative Results

We finally present some qualitative results in Figure 4. In most cases the network predicts actions correctly. One error is made in the second image, where a left turn is incorrectly predicted as possible. This is likely due to the fact that it is hard to infer left or right turns in the middle of the crossroad. It is also safe to say that the network can successfully pick the few objects that are action-inducing, including small traffic signs, lights, or obstacles on the side of the road, while ignoring many other objects that are not action-inducing. This is unlike the method of [5], whose selector fails to capture most action-inducing objects, leading to more errors in explanation prediction.

6. Conclusion

In this work, we propose the problem of object induced action and explanation prediction for autonomous driving. A dataset was introduced for this task and a new architecture proposed for its solution. The new task is an interesting challenge for computer vision, because it requires object reasoning that accounts for scene context. The goal is not simply to detect objects, but to detect objects that could create hazard in the autonomous driving setting, and produce explanations for all actions predicted. However, because all explanations are grounded in objects that induce actions, they are easier to generate than in the generic computer vision setting. In fact, they reduce to one-hot style prediction and can be addressed with classification techniques. Due to this, the addition of explanations was shown to increase the accuracy of action predictions in our experiments. We believe that this is the first showing of explanations actually helping improve the performance of a deep learning system.

Acknowledgements

This work was partially funded by NSF awards IIS-1637941, IIS-1924937, and NVIDIA GPU donations.

References

- [1] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2942–2950, 2017.
- [2] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2174–2182, 2017.
- [3] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.
- [4] Dequan Wang, Coline Devin, Qi-Zhi Cai, Philipp Krähenbühl, and Trevor Darrell. Monocular plan view networks for autonomous driving. *arXiv preprint arXiv:1905.06937*, 2019.
- [5] Dequan Wang, Coline Devin, Qi-Zhi Cai, Fisher Yu, and Trevor Darrell. Deep object-centric policies for autonomous driving. In *International Conference on Robotics and Automation (ICRA)*, pages 8853–8859. IEEE, 2019.
- [6] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [7] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11867–11876, 2019.
- [8] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7644–7652, 2019.
- [9] Gregory P Meyer, Ankit Laddha, Eric Kee, Carlos Vallespi-Gonzalez, and Carl K Wellington. Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12677–12686, 2019.
- [10] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8453, 2019.
- [11] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7026–7035, 2019.
- [12] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10986–10995, 2019.
- [13] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12240–12249, 2019.
- [14] Chen-Hsuan Lin, Oliver Wang, Bryan C Russell, Eli Shechtman, Vladimir G Kim, Matthew Fisher, and Simon Lucey. Photometric mesh optimization for video-aligned 3d object reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 969–978, 2019.
- [15] Car Crash. Enabling pedestrian safety using computer vision techniques: A case study of the 2018 uber inc. self-driving. In *Advances in Information and Communication: Proceedings of the 2019 Future of Information and Communication Conference (FICC)*, volume 1, page 261. Springer, 2019.
- [16] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260, 2017.
- [17] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279, 2018.
- [18] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian Conference on Computer Vision*, pages 658–674. Springer, 2018.
- [19] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver’s focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018.
- [20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012.
- [21] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.
- [22] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [23] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Buló, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

- [24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [25] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision (ECCV)*, pages 818–833, 2014.
- [26] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in neural information processing systems*, pages 1495–1503, 2015.
- [27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.
- [28] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018.
- [29] Dean Pomerleau. Alvin: An autonomous land vehicle in a neural network. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems 1*. Morgan Kaufmann, January 1989.
- [30] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Praseoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016.
- [31] Bogdan Alexe, Nicolas Heess, Yee W Teh, and Vittorio Ferrari. Searching for objects driven by context. In *Advances in Neural Information Processing Systems*, pages 881–889, 2012.
- [32] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.
- [33] Coline Devin, Pieter Abbeel, Trevor Darrell, and Sergey Levine. Deep object-centric representations for generalizable robot learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 7111–7118, 2018.
- [34] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016.
- [35] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 322–338, 2018.
- [36] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–685, 2018.
- [37] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018.
- [38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [39] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [40] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.
- [41] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry D. Jackel, Urs Muller, and Karol Zieba. Visualbackprop: visualizing cnns for autonomous driving. *CoRR*, abs/1611.05418, 2016.
- [42] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep Learning. *Nature*, 521(7553):436–444, May 2015.
- [43] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.
- [44] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.