CVPR JUNE 14-19 2020

Explainable Object-induced Action Decision for Autonomous Vehicles

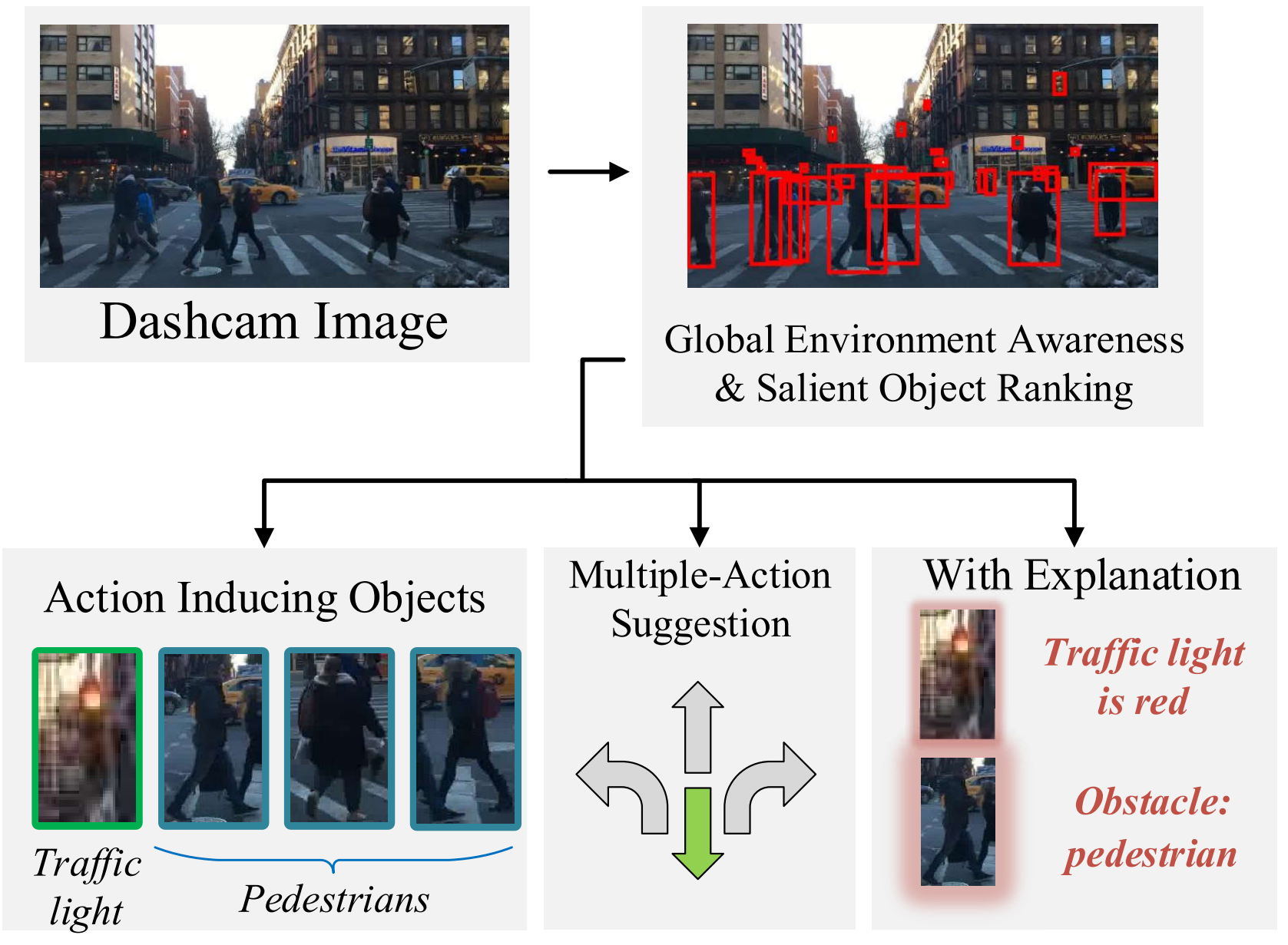
Yiran Xu Xiaoyin Yang Lihang Gong Hsuan-Chu Lin Tz-Ying Wu Yusheng Li Nuno Vasconcelos
UC San Diego

SVCL

I. Introduction

- Motivation

We propose a new paradigm for autonomous driving. The new paradigm lies between end-to-end and pipelined approaches and it inspired by how humans solve the problem. We denote this new paradigm as action-inducing since changes in the object states can trigger vehicle actions. Also, a set of explanations for actions are introduced.



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.

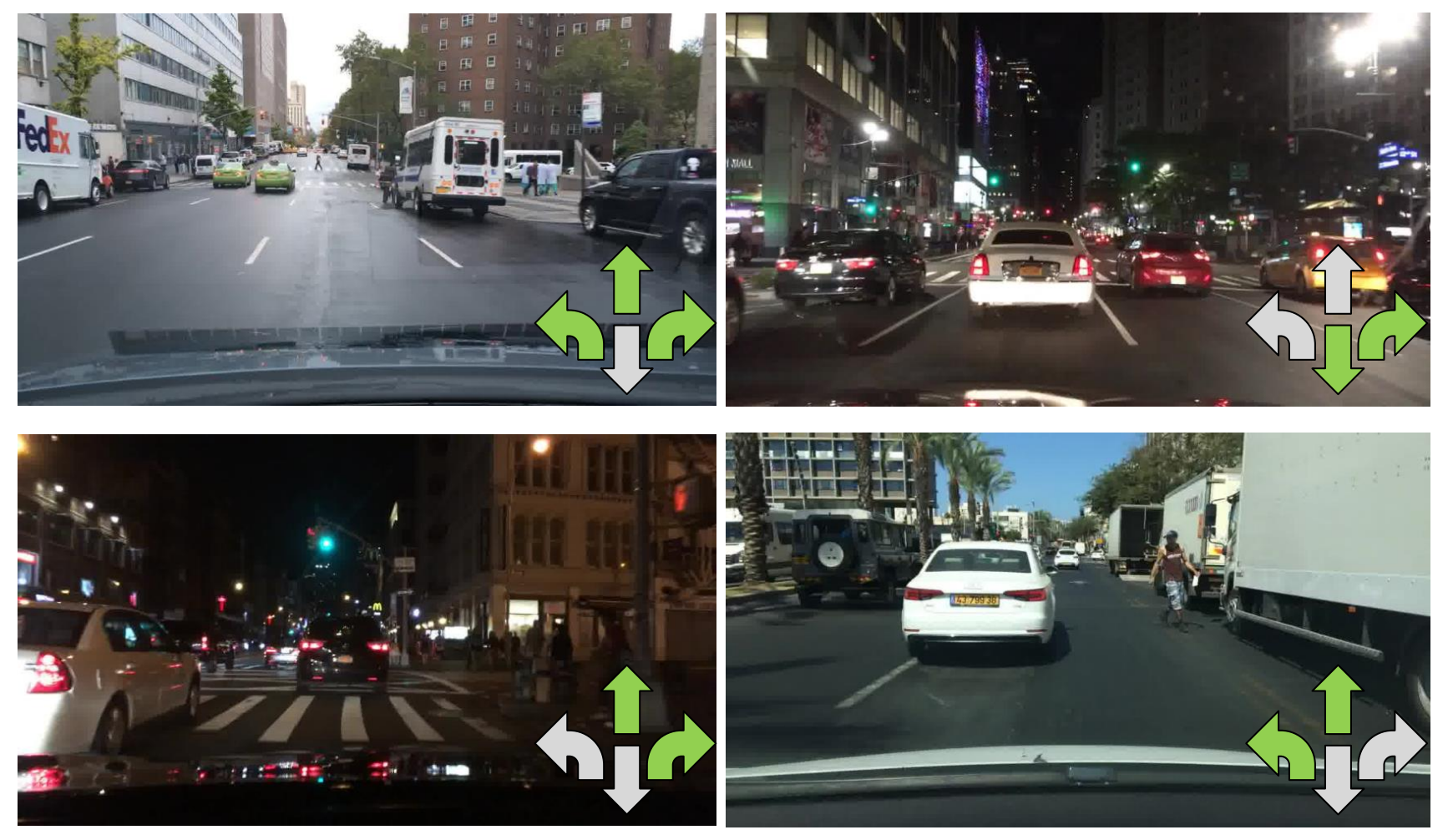
- Contribution
  - A **large dataset** annotated for both driving commands and explanations
  - A **new multi-task formulation** of the action prediction problem
  - A CNN architecture as the solution
  - An experimental evaluation showing the explanations improve the decision making for actions

II. Object-Inducing Learning

- Object-Inducing Action Prediction with Explanation

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

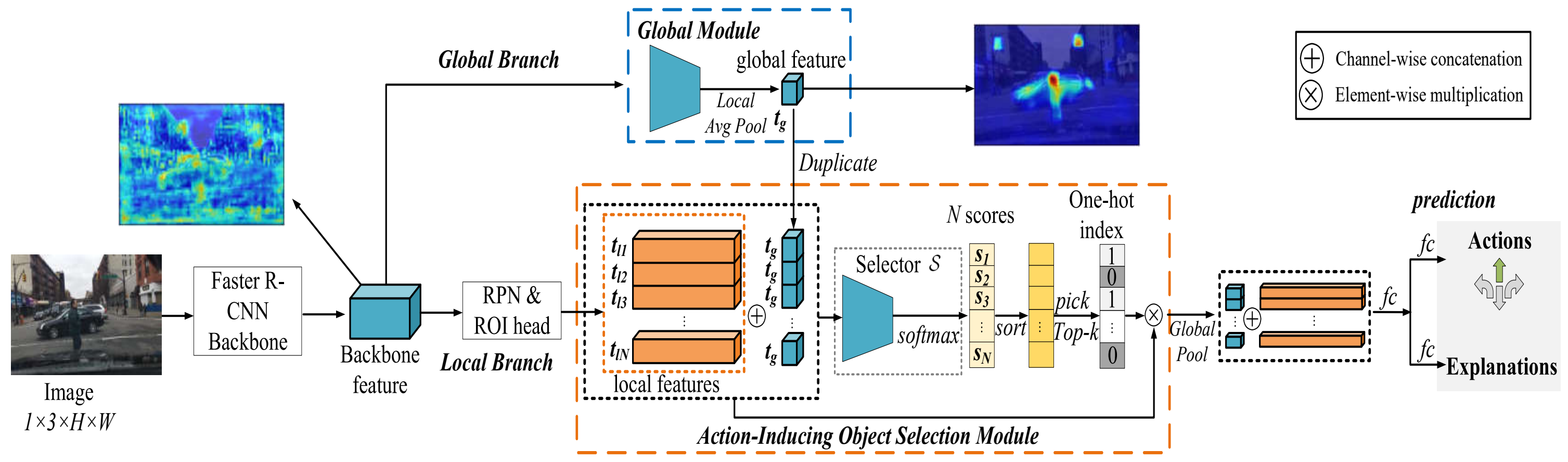
- BDD-OIA Dataset
  - An extension of BDD100K dataset with newly annotated actions and explanations for each scenario.



  - Downloadable here.

III. Method

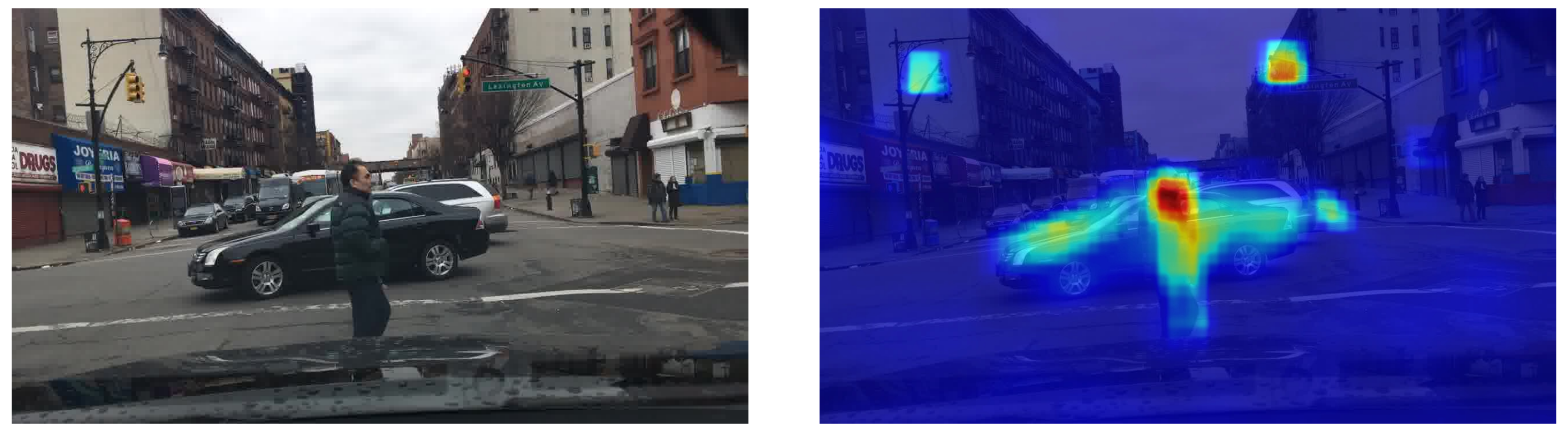
- Overall Architecture
  - The Faster R-CNN is used to extract backbone features, which are fed into a global and a local branch. The Global Module generates a global feature map that provides scene context, while the local branch captures the details of action-inducing objects. In the local branch, a selector module outputs a score for each object feature tensor and associated global context information. The top kaction-inducing objects are selected and the features from the two branches are concatenated for action and explanation prediction.



- Multi-task Learning
  - $\mathcal{L}_A$ is the loss of action prediction and $\mathcal{L}_E$ is the loss of explanation. λ represents the strength of explanation.

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

- 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.
  - **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} 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. A selector $\mathcal{S}$ then chooses the action-inducing objects from this tensor.
  - **Behind these.** Two visualizations derived from the input image are shown below. Heatmap is the output from the global module. The combination of local and global features and end-to-end supervision leads 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.



- Evaluation Metrics

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

$$\mathrm{mF1} = \mathrm{mean}(F1_j)\,. \qquad (4)$$

IV. Experiment

- Interplay between Actions and Explanations
  - 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). "A" stands for action and "X" stands for explanation.

λ	F	S	L	R	A mF1	A $\mathrm{F1}_{all}$	X $\mathrm{F1}_{all}$
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

- Interplay between Local and Global Features

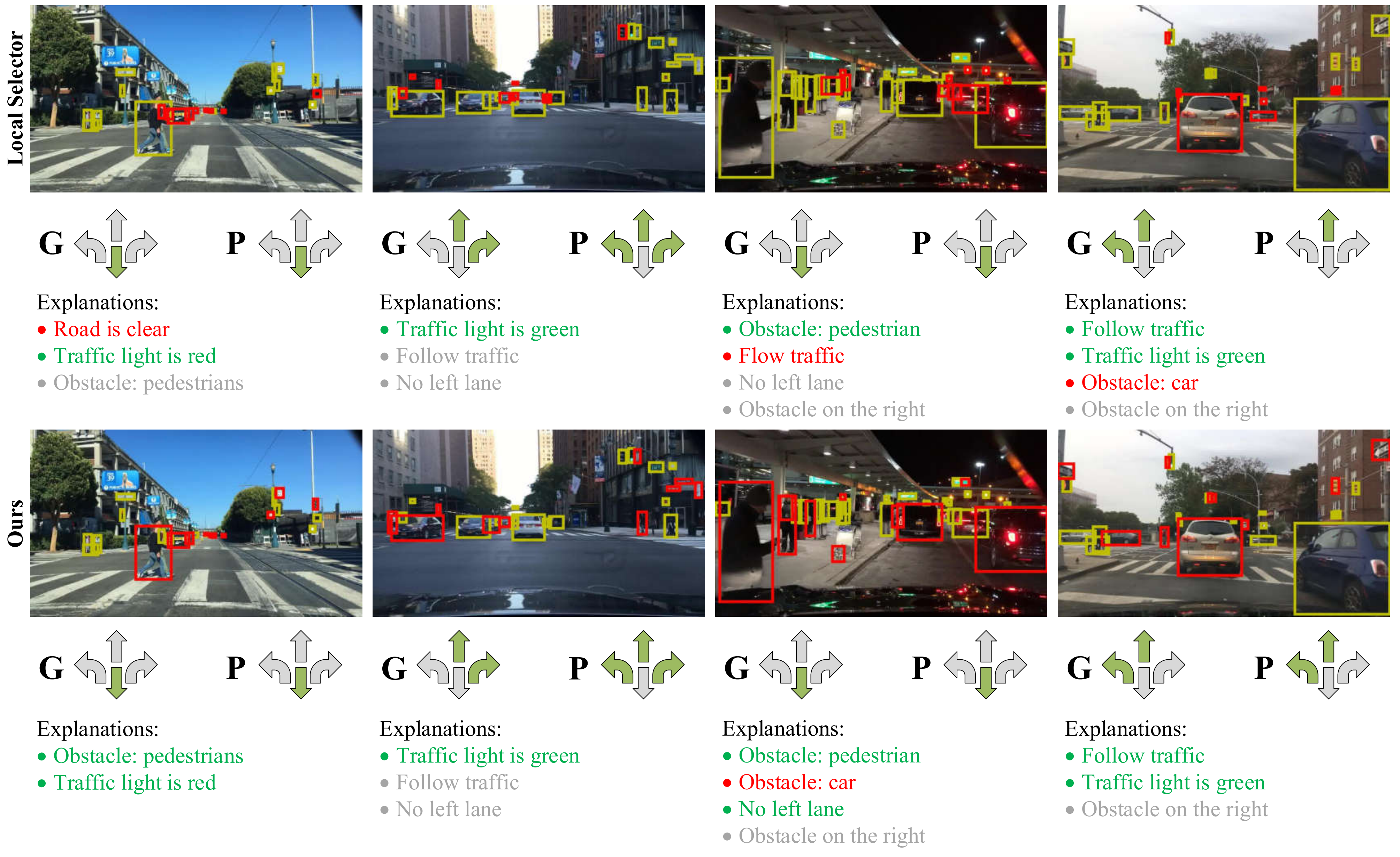
models	F	S	L	R	A mF1	A $\mathrm{F1}_{all}$	X mF1	X $\mathrm{F1}_{all}$
only local branch	0.760	0.649	0.413	0.473	0.574	0.605	0.139	0.351
only global branch	0.820	0.777	0.499	0.621	0.679	0.704	0.206	0.419
random selection	0.823	0.778	0.499	0.637	0.685	0.709	0.197	0.413
select top-5	0.821	0.768	0.617	0.625	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**

- Model Comparisons

models	F	S	L	R	mF1	$\mathrm{F1}_{all}$	explanation mF1	explanation $\mathrm{F1}_{all}$
Baseline	0.755	0.607	0.098	0.108	0.392	0.601	0.180	0.331
local selector	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**

V. Visulization

- 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).



Explanations:
- Road is clear
- Traffic light is red
- Obstacle: pedestrians

Explanations:
- Traffic light is green
- Follow traffic
- No left lane

Explanations:
- Obstacle: pedestrian
- Flow traffic
- No left lane
- Obstacle on the right

Explanations:
- Follow traffic
- Traffic light is green
- Obstacle: car
- Obstacle on the right

Explanations:
- Obstacle: pedestrians
- Traffic light is red

Explanations:
- Traffic light is green
- Follow traffic
- No left lane

Explanations:
- Obstacle: pedestrian
- Obstacle: car
- No left lane
- Obstacle on the right

Explanations:
- Follow traffic
- Traffic light is green
- Obstacle on the right

CVPR, Seattle, United States, 2020