RS2G: Data-Driven Scene-Graph Extraction and Embedding for Robust Autonomous Perception and Scenario Understanding

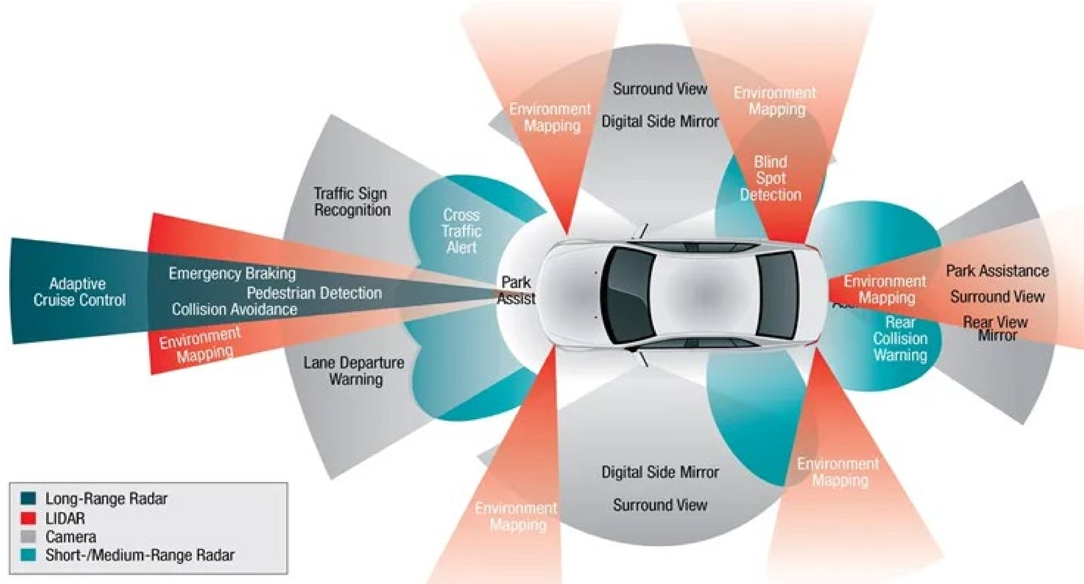
Junyao Wang, Arnav Vaibhav Malawade, Junhong Zhou, Shih-Yuan Yu, Mohammad Abdullah Al Faruque
{junyaow4, malawade, junhonz2, alfaruqu}@uci.edu, University of California, Irvine, United States

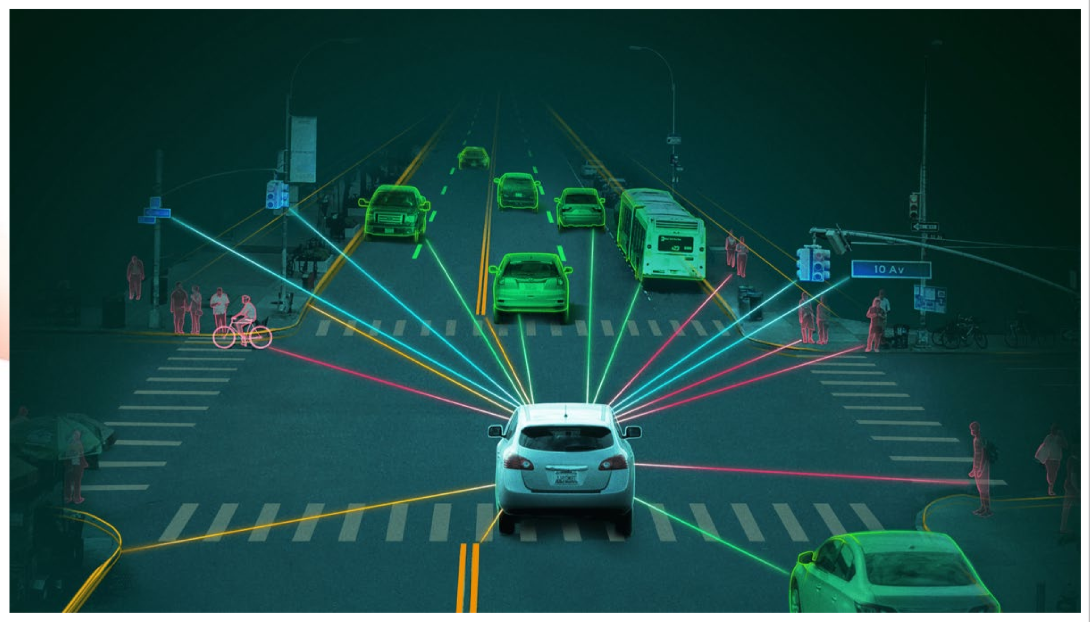
Motivation

- RS2G focuses on **subject risk assessment**.
- Effectively modeling the **relations** among road users is very important for autonomous vehicle to understand the surrounding environment.

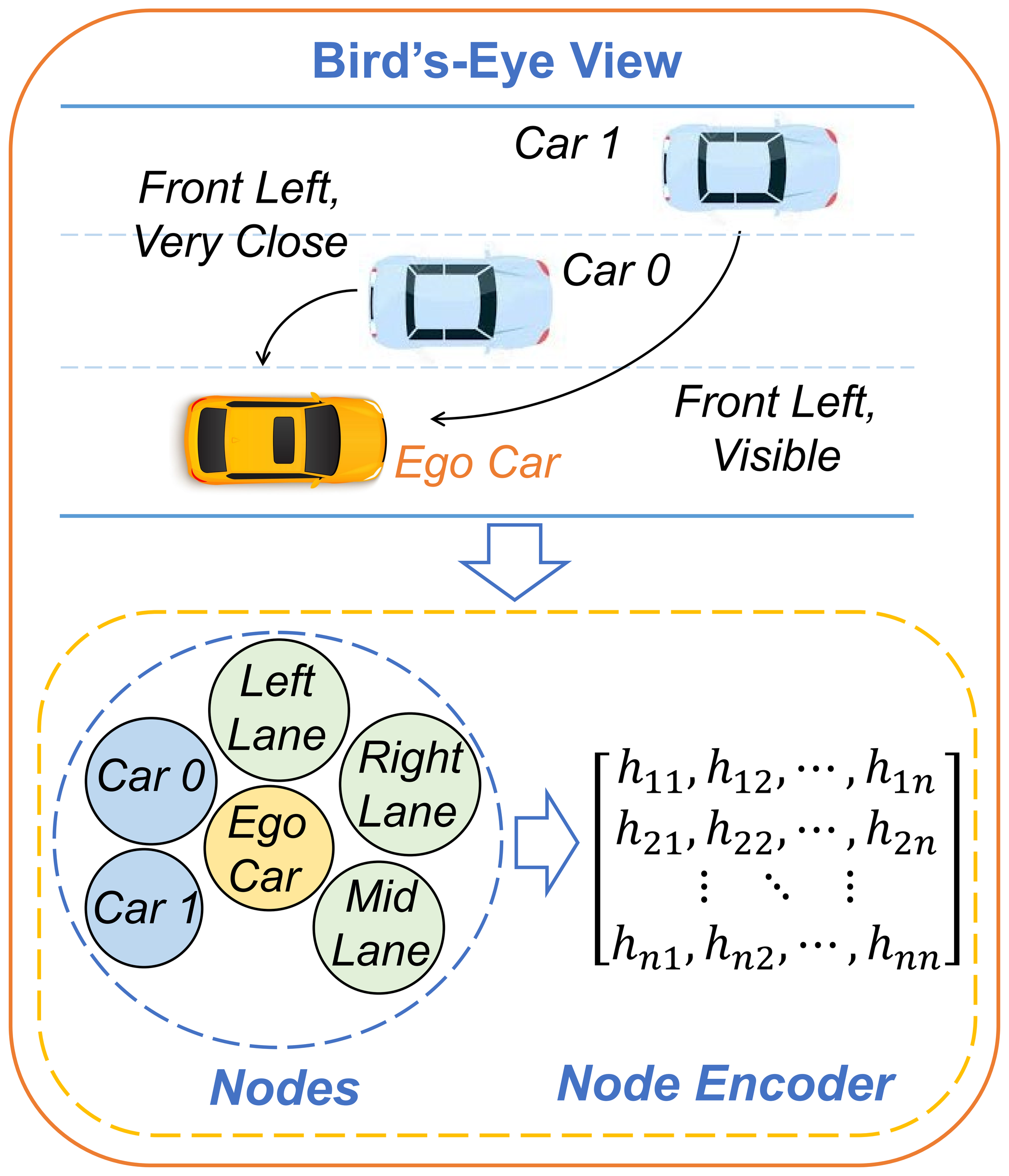
Challenges:

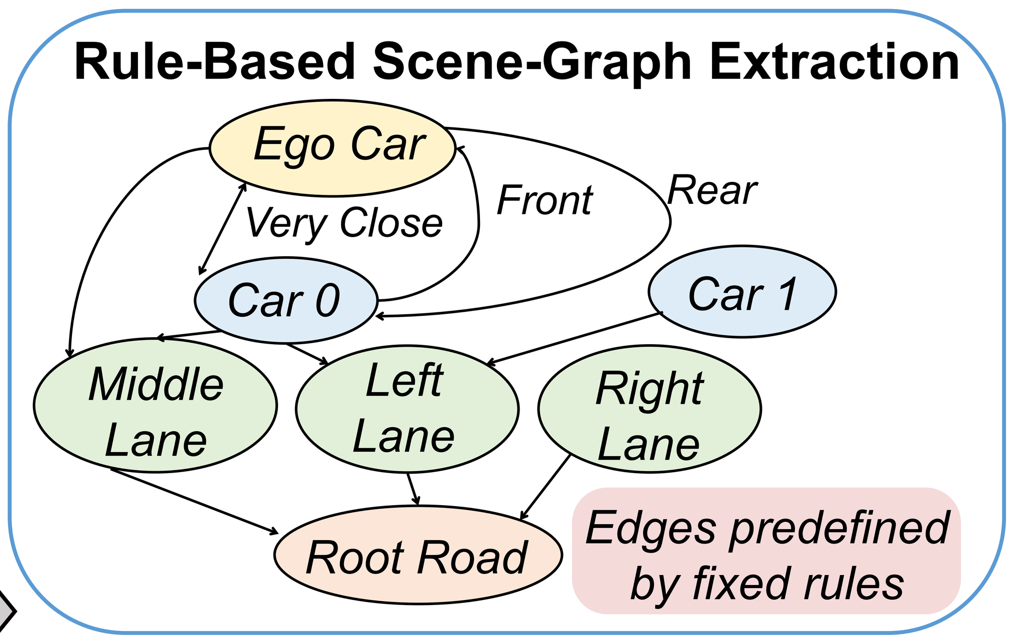
- Convolution Neural Networks (CNNs) often fail to account for high-level semantic scenes (rarely consider ***interactions*** between driving agents and environmental factors)
- Existing graph learning (GL) models rely on predefined domain-specific graph extraction rules ⟹ severely impedes model ***generalizability***.

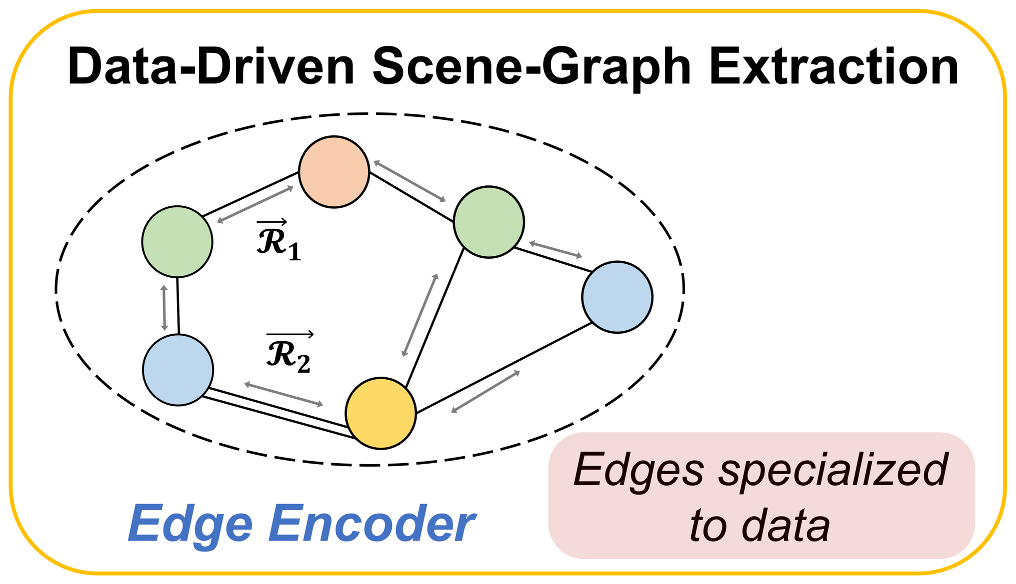




Objective: Data-Driven Scene-Graph Representation



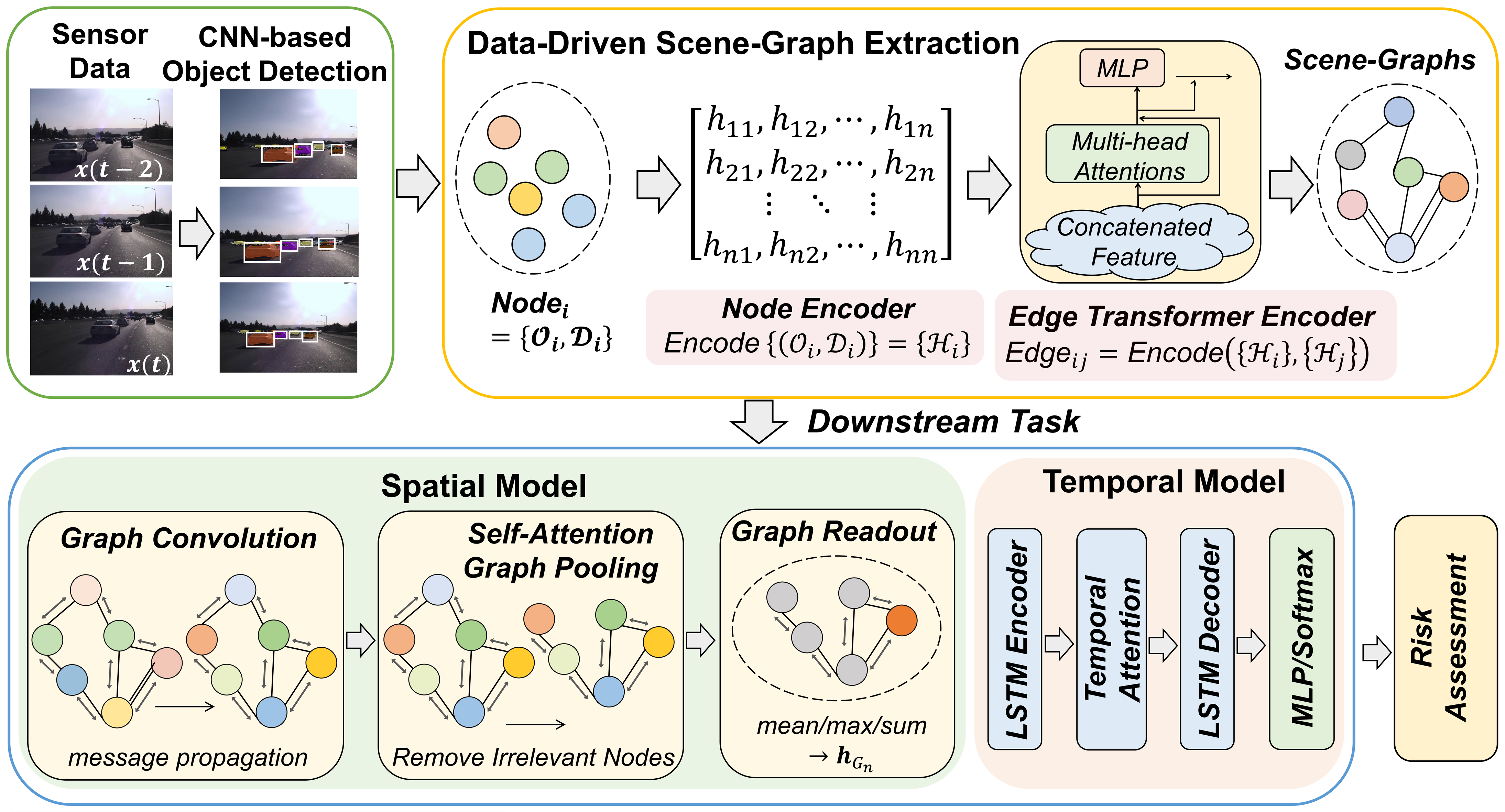




- Both rule-based and data-driven Scene-Graph extraction methods start with transforming objects to nodes with a node encoder.
- The rule-based scene-graph extraction relies on ***fixed rules*** derived from expert knowledge; its encoded edges typically have ***concrete physical meanings*** and the graphs are ***constrained by specific domains***.
- ***Our data-driven scene-graph extraction represents diverse relations between nodes with vectors, which better captures latent features and can be more dynamic and domain-adaptive.***

Methodology

The Architecture of Our Proposed RS2G:



Algorithm 1: Data-Driven Scene-Graph Extraction

Input: Objects $\mathcal{O}_t$ and their attributes $\mathcal{D}_t$ at time t.
Output: Scene-graph $\mathcal{G}_t$ at time t.
def $\Psi\,(\mathcal{O}_t, \mathcal{D}_t)$ **:**
  $\mathcal{H}_t \leftarrow \emptyset$, $\mathcal{A}_t \leftarrow \mathbf{0}_{n\times n}$ ▷ initialize outputs
  for $o_j, \mathbf{d}_j \in \mathcal{O}_t, \mathcal{D}_t$ **do**
    $\mathbf{h}_j \leftarrow Encode_{node}(o_j, \mathbf{d}_j)$ ▷ node encoding
    $\mathcal{H}_t.append(\mathbf{h}_j)$
  $\mathcal{C} \leftarrow \mathcal{H}_t \times \mathcal{H}_t$ ▷ get all pair of nodes
  for *relation* $r \in \mathcal{R}$ **do**
    for *edge* $(\mathbf{h}_j, \mathbf{h}_k) \in \mathcal{C}$ **do**
      $(\mathcal{A}_t)_{r,j,k} \leftarrow MLP(Encode_{edge}(r, \mathbf{h}_j, \mathbf{h}_k))$
  $\mathcal{G}_t \leftarrow \{\mathcal{H}_t, \mathcal{A}_t\}$
  return $\mathcal{G}_t$

- RS2G starts with a set of objects and their attributes extracted by a pre-trained CNN-based model
- We then utilize our data-driven scene-graph extraction to generate a set of scene-graphs of the current scene (Algorithm 1)
- We analyze scene-graphs with our spatial-temporal embedding model, consisting of a multi-relational graph convolutional network (MR-GCN) and a long short-term memory (LSTM) network
- Finally, we utilize a multi-layer perceptron (MLP) to classify the risk of the driving scenario as risky or non-risky

Selected Experimental Result

Subjective Risk Assessment

Dataset	Graph Extraction	Accuracy	MCC	AUC
271-carla	None	73.17%	0.1887	0.8043
	Rule-Based	82.93%	0.5173	0.8098
	RS2G (1D MLP)	84.51%	0.2093	0.9338
	RS2G (2D MLP)	**86.59%**	**0.468**	**0.9578**
	RS2G (Transformer)	84.15%	0.402	0.9362
1043-carla	None	71.66%	0.1111	0.7173
	Rule-Based	91.43%	0.7217	0.971
	RS2G (1D MLP)	91.72%	0.6840	0.9643
	RS2G (2D MLP)	93.31%	0.7426	0.7949
	RS2G (Transformer)	**97.13%**	**0.8823**	**0.9686**
1361-honda	None	60.39%	0.0391	0.7110
	Rule-Based	86.31%	0.2445	0.9341
	RS2G (1D MLP)	87.04%	0.1626	0.9315
	RS2G (2D MLP)	89.00%	0.3029	0.9383
	RS2G (Transformer)	**89.98%**	**0.404**	**0.9495**
620-dash	None	48.92%	-0.1749	0.5256
	Rule-Based	67.20%	0.3428	0.6966
	RS2G (1D MLP)	68.82%	0.3967	0.7403
	RS2G (2D MLP)	**72.04%**	**0.4398**	**0.8047**
	RS2G (Transformer)	68.28%	0.3635	0.7354

Transfer Learning

Dataset	Graph Extraction	Accuracy	MCC	AUC
271-carla to 620-dash	None	52.58%	0.0333	0.5126
	Rule-Based	48.22%	0.0238	0.4975
	RS2G(2D MLP)	57.25%	0.1398	0.5669
	RS2G(Transformer)	**64.68%**	**0.2957**	**0.6831**
1043-carla to 620-dash	None	49.03%	-0.0432	0.4999
	Rule-Based	50.96%	0.0021	0.5093
	RS2G(2D MLP)	60.65%	0.2089	0.6265
	RS2G (Transformer)	**66.29%**	**0.3293**	**0.6964**

Ablation Studies

Impacts of Downstream Components

Graph Extraction	Spatial Model	Temporal Model	Accuracy	MCC	AUC
Rule-Based	MLP	mean	52.15%	0.0000	0.4973
Rule-Based	MLP	LSTM	62.90%	0.2741	0.6811
Rule-Based	MRGCN	mean	63.44%	0.2696	0.6867
Rule-Based	MRGCN	LSTM	**75.27%**	**0.5197**	**0.8248**
RS2G	MLP	mean	81.74%	0.1857	0.9228
RS2G	MLP	LSTM	81.45%	0.402	0.9472
RS2G	MRGCN	mean	**87.80%**	**0.5403**	**0.9468**
RS2G	MRGCN	LSTM	84.15%	0.402	0.9362

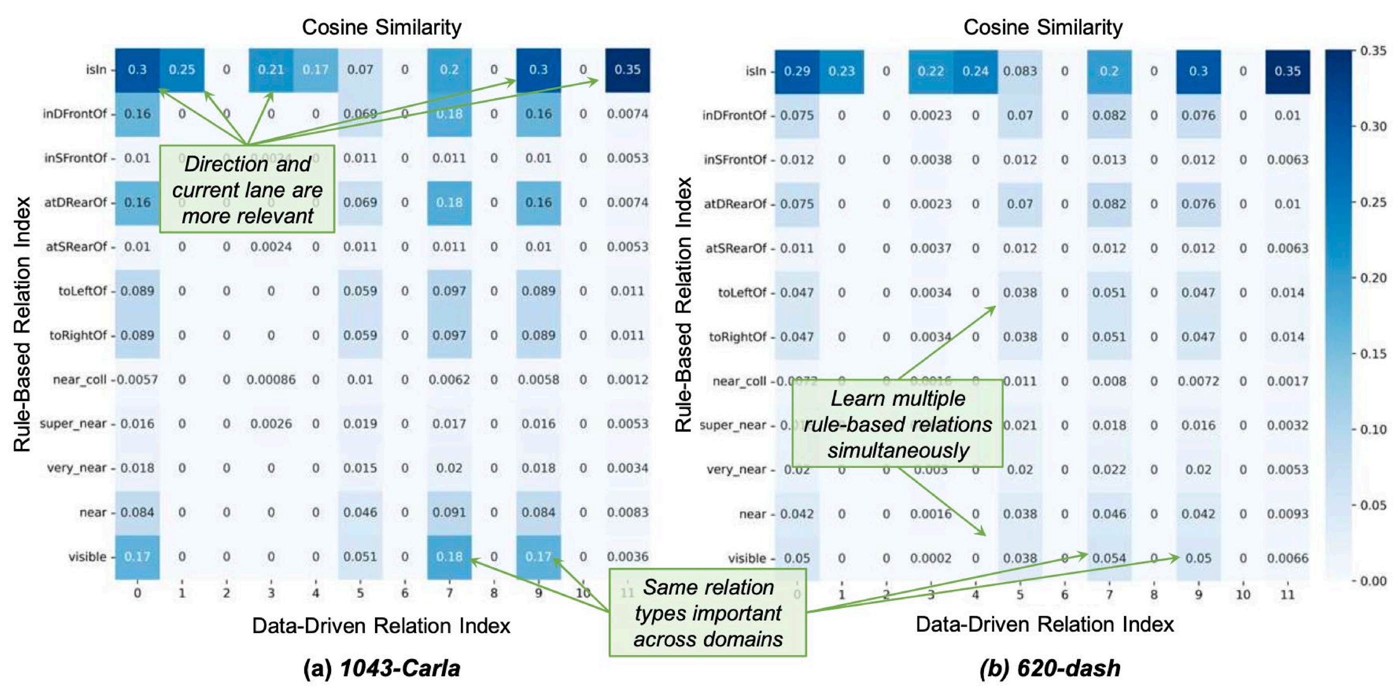
Impacts of KL Divergence

Model		Accuracy	MCC	AUC
Graph Extraction	with KL			
RS2G(2MLP)	✗	62.42%	0.2455	0.6132
RS2G(2MLP)	✓	60.65%	0.2089	0.6265
RS2G(Transformer)	✗	64.35%	0.2897	0.6586
RS2G(Transformer)	✓	**66.29%**	**0.3293**	**0.6964**

Impacts of Edge Extraction Threshold

Graph Ext.	Acc.	Avg. Deg.	Avg. Edges	σ Edges
Rule-Based	95.86%	3.84	16.50	10.51
RS2G ($\gamma = 0.25$)	**97.13%**	**37.11**	**298.98**	**264.36**
RS2G ($\gamma = 0.5$)	94.59%	23.98	193.21	171.22
RS2G ($\gamma = 0.75$)	95.54%	10.88	87.68	78.00

Cosine Relation Similarity



(a) 1043-Carla (b) 620-dash

Selected References:
[1] Shih-Yuan Yu, et al.,Scene-graph augmented data0driven risk assessment of autonomous vehicles decisions. IEEE Transactions on Intelligent Transportation Systems, 2021
[2] Ekim Yurtsever, et al., Risky actions recognition in lane change video clips using deep spatiotemporal networks with segmentation mask transfer. IEEE Intellignet Transportation Systems Conference (ITSC), 2019.