Supplementary Materials
Learning Road Scene-level Representations via Semantic Region Prediction

Zihao Xiao, Alan Yuille
Department of Computer Science
Johns Hopkins University
{zxiao10,ayuille1}@jhu.edu

Yi-Ting Chen
Department of Computer Science
National Yang Ming Chiao Tung University
ychen@nycu.edu.tw

1 Automatic Semantic Region Labeling

We use a *Right-turn* sample to illustrate the automatic semantic region labeling process. We derive a semantic BEV image with the process described in the main paper. As shown in Figure 1, camera locations overlapping with the first and second crosswalks are annotated as A_i and C_i, respectively. The poses locating between A_i and C_i are annotated as B_i. Camera poses locating in areas before the first and the second *crosswalk* are S_i and T_i, respectively. Note that the parameter i is 3 because this is a *Right-turn* at a 4-way intersection. The parameter i is set to 1 and 2 for *Left-turn* and *Go Straight*, respectively.

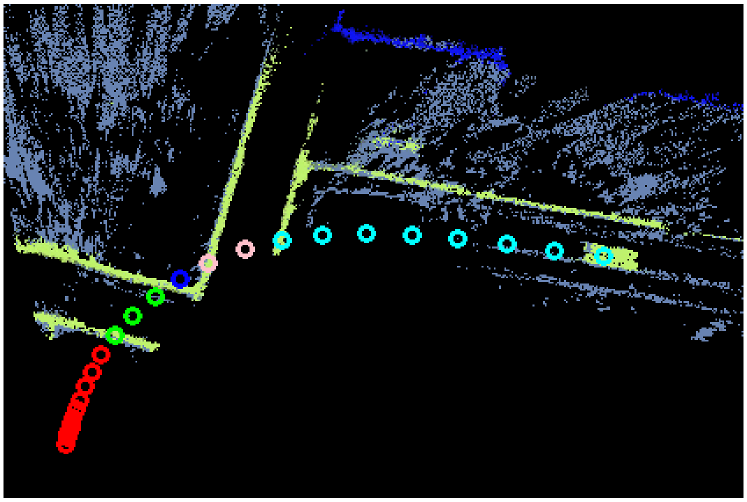


Figure 1: **Sample of automatic labeling.** We show the results of automatic semantic region labeling of a *Right-turn* sample. The semantic regions are visualized in colors. The red circle indicates S. The green circle indicates A_3. The blue circle indicates B_3. The pink circle indicates C_3. The cyan circle indicates T_3. Best Viewed in color.

To evaluate the effectiveness of our automatic semantic region labeling process, we randomly pick 100 video clips and annotate ground truth semantic regions manually. The accuracy of automatic semantic labeling is 76.4%. We diagnose the results and find the following reasons for failures. First, some video clips do not start from the semantic region S because the original starting time labeled in the HDD dataset is inaccurate. Second, lines like lane-changing lines and arrows indicating directions are wrongly predicted as crosswalks by the segmentation model. To improve the quality of labeling, we plan to annotate the center of the 4-way intersection (i.e., B_i) and train another semantic segmentator to mitigate the second issue in future work.

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

Model	Feature	HDD			HDD Interactive		
		Macro Avg Pre	Micro Avg Pre	mAP	Macro Avg Pre	Micro Avg Pre	mAP
SRP-INT	ImageNet	51.3	**74.6**	53.9	45.3	59.3	60.1
SRP-INT	nuScenes	25.0	57.8	45.5	**67.4**	69.8	69.1
SRP-INT	COCO Panoptic	52.8	61.8	51.7	46.1	61.1	63.3
SRP-INT	Mapillary Vistas	**55.3**	73.8	**57.9**	67.0	**70.3**	**69.5**

Table 1: **Ablation study for model pretraining in HDD dataset on driver intention prediction.** Base model pretrained on the Mapillary Vistas dataset leads to better performance in general. The results confirm the importance of the final task of a pretraining model.

Model	Feature	nuScenes		
		Macro Avg Pre	Micro Avg Pre	mAP
SRP-INT	ImageNet	36.0	59.7	58.1
SRP-INT	nuScenes	**45.1**	37.6	59.5
SRP-INT	COCO Panoptic	37.6	63.8	61.1
SRP-INT	Mapillary Vistas	41.1	**68.3**	**66.7**

Table 2: **Ablation study for model pretraining in nuScenes dataset on driver intention prediction.** Base model pretrained on the Mapillary Vistas dataset leads to better performance in general. The results confirm the importance of the final task of a pretraining model.

2 Experimental Details

Driver Intention Prediction. After training SRP, we freeze every other layers but the intention classifier. Similar to training SRP, We use Adam optimizer [1] with default parameters, a learning rate of 0.0001, and weight decay of 0.0005. The model is trained for 60 epochs. We report the performances of the last epoch.

Risk Object Identification. We make use of the same weights of SRR as SRP-INT does and the weights are frozen during the training process. The hidden state of the SRP is connected to a fully connected(FC) layer before being fussed with the ego representation. We then follow the two-stage strategy as described in the main paper to obtain the risk object predictions.

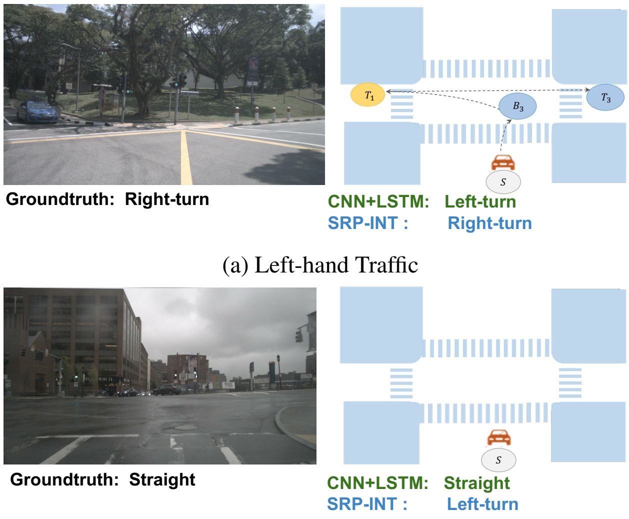
3 Ablation Study: Model Pretraining

We evaluate the impact of model pretraining for the base model. We follow the same training procedure as the main paper and diagnose our SRP-INT on the HDD, HDD interactive and nuScenes.

As shown in Table 1 and Table 2, the backbone model pretrained on the Mapillary Vistas dataset results in significantly better performance compared with the backbone models trained on other datasets. Note that these backbones are trained on different tasks. The Mapillary Vistas backbone is pre-trained on the panoptic segmentation task. The nuScenes backbone is trained on instance segmentation task using data released in nuImage, an extension of nuScenes that contains additional images and 2D annotations. Note that they have semantic labels for the drivable surface. The COCO Panoptic backbone is trained on COCO Panoptic Segmentation. The model performs favorably in different settings, while COCO Panoptic is not a traffic scene dataset. The tables show that the in-domain nuScenes backbone cannot perform well in many metrics. We hypothesize that the two tasks, i.e., intention prediction and risk object identification, require *Stuff* information (e.g., road, lane marking, and crosswalk). On the other hand, the nuScenes backbone learns to detect objects, which could explain the superior performance on HDD interactive cases because these cases involve interaction with other traffic participants.

4 Generalization to nuScenes

In the task of driver intention prediction, to demonstrate the effectiveness of the learned representations, we first train our model on the HDD dataset [2] and test on the nuScenes dataset [3] without finetuning. It is worth noting that the domain gap between the HDD dataset and the nuScenes dataset is significant and could lead to false predictions in either semantic region predictions or intention predictions. Some videos in the nuScenes dataset are collected in countries with left-hand traffic, while data in HDD dataset is in right-hand traffic conditions. Another typical failure occurs when the ego vehicle approaches an empty 4-way intersection. As shown in Fig. 2b, it is challenging to make correct predictions without additional cues. One possible future direction is to leverage drivers' gazes as in the Brain4Car project [4] or steering signals.



(a) Left-hand Traffic

(b) Empty Intersection

Figure 2: **Failure cases on the nuScenes dataset.** We show two typical failure cases of SRP-INT on the nuScenes dataset. We provide ground truth as well as the driver intention predictions of the CNN+LSTM baseline and our proposed SRP-INT. The semantic region predictions are shown on the right side.

5 Interactive Scenarios

It is challenging to predict the driver's intention in the modality of monocular image sequences due to complicated driving scenarios such as drivers may have to stop for crossing vehicles or yield to crossing pedestrians before they reach their intended goals. We call these cases interactive scenarios. We evaluate our model on interactive scenarios on the HDD testing set because of their importance in real-world applications. We present quantitative and qualitative evaluations of interactive scenarios in the main paper as well as the following sections of the supplementary materials.

6 Qualitative Results

We show qualitative results of driver intention prediction on the HDD dataset and nuScenes dataset of SRP-INT in Fig. 3 and Fig. 4, respectively. For comparison, we also show the intention predictions of the CNN+LSTM baselines as the intention ground truth.

Qualitative results of SRP-ROI on two risk object categories: Crossing Vehicle and Crossing Pedestrian are shown in Fig. 5 and Fig. 6, respectively.

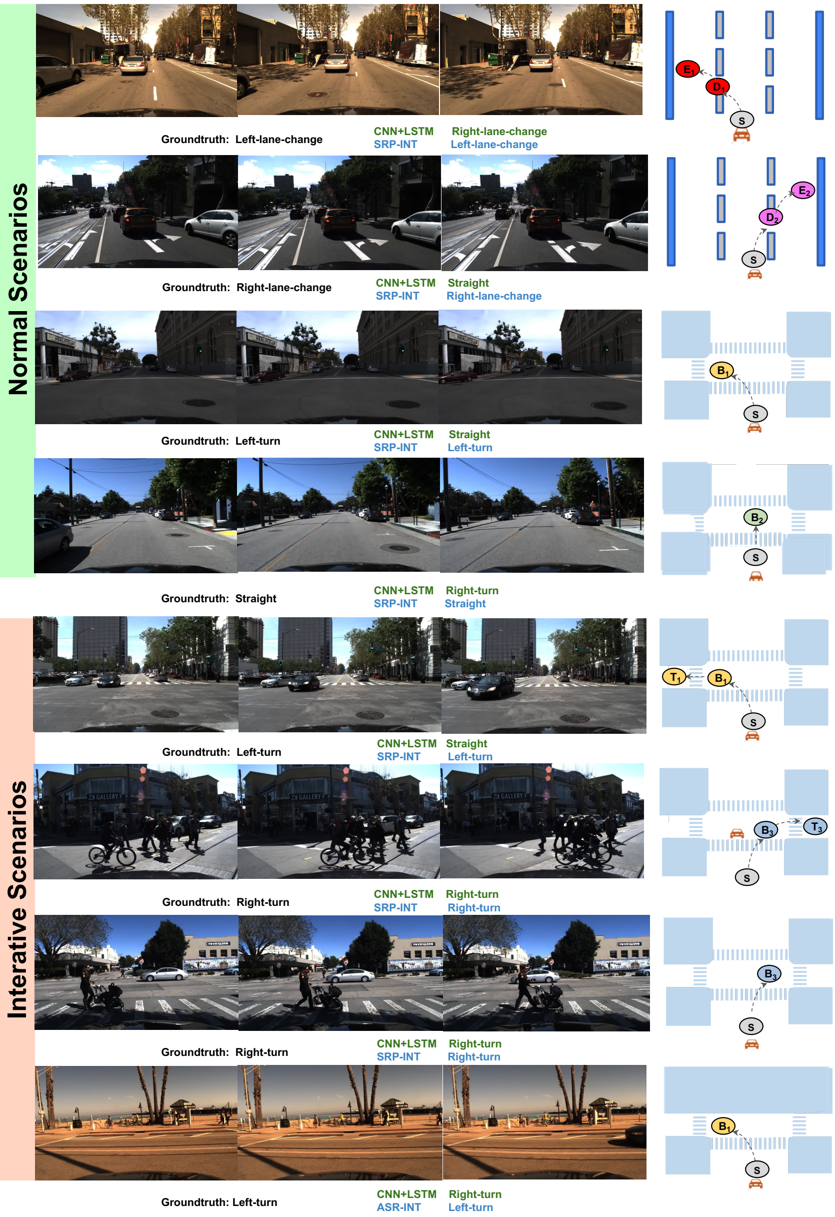


Figure 3: **Qualitative results of driver intention prediction on the HDD dataset [2].** The examples shown in the first four rows indicate normal cases, i.e., the ego-vehicle navigating through the intersection without interactions with other traffic participants. The examples shown in the last four rows are interactive scenarios. We provide the ground truth of ego-vehicle intention and the prediction of the final SRP-INT and the CNN+LSTM baseline on the HDD dataset. The predictions of semantic regions of SRP-INT are displayed on the right side of each scenario, where traffic participants intervene in the movement of the ego-vehicle. The results demonstrate the proposed framework can predict semantic regions reliably, and that helps the visual system predict ego-vehicle intention. The qualitative experiments empirically justify the value of the proposed scene-level representation learning.

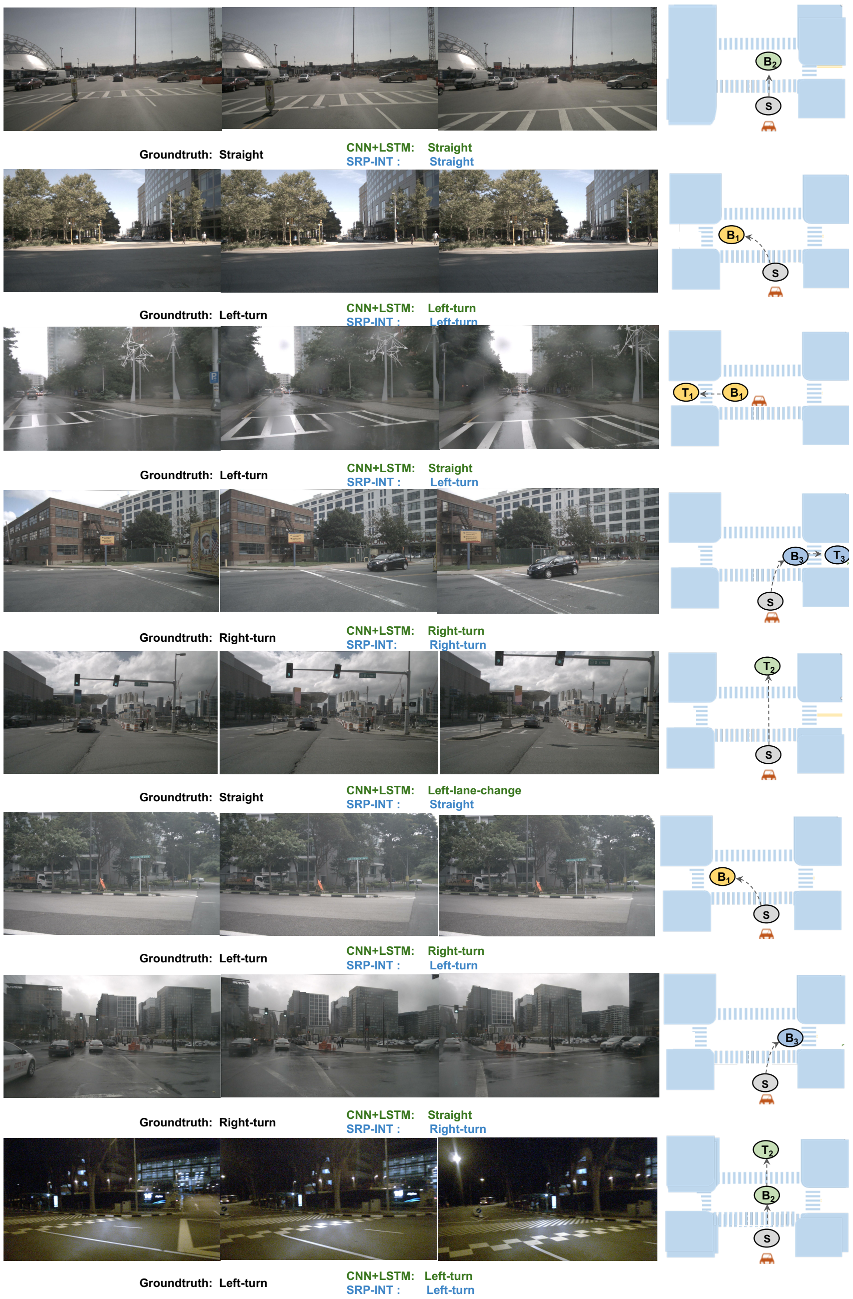


Figure 4: **Qualitative results of driver intention prediction on the nuScenes dataset [3].** Similar to the results on the HDD dataset [2] shown in Figure 3, the ground truth of ego-vehicle intention, as well as the prediction of the SRP-INT and the CNN+LSTM baseline on the nuScenes dataset, are presented. The semantic region predictions are provided on the right side of each case.

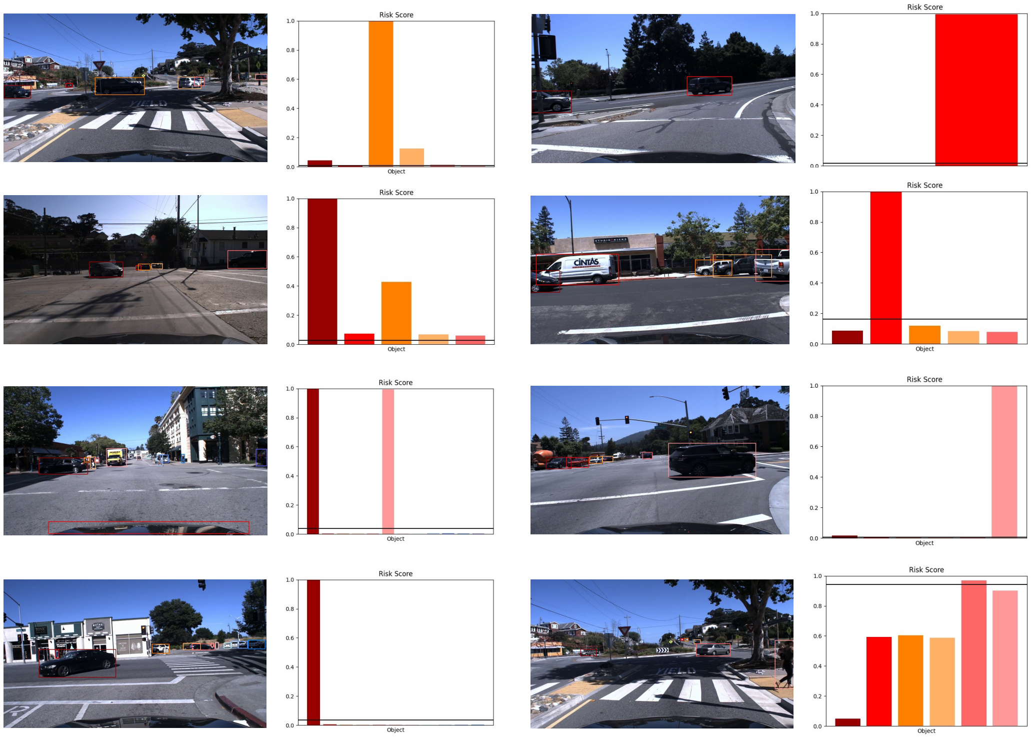


Figure 5: **Qualitative results of risk object identification – Crossing-Vehicle.** We demonstrate the effectiveness of the proposed scene-level representation for risk object identification. According to the definition of the risky object proposed in [5], the candidate with the highest risk score is the risk object. In this figure, we show the risk scores of each object candidate and demonstrate the system can differentiate risk and non-risk objects in various crossing vehicle scenarios. For each candidate, the color of the bar matches the color of the bounding box.

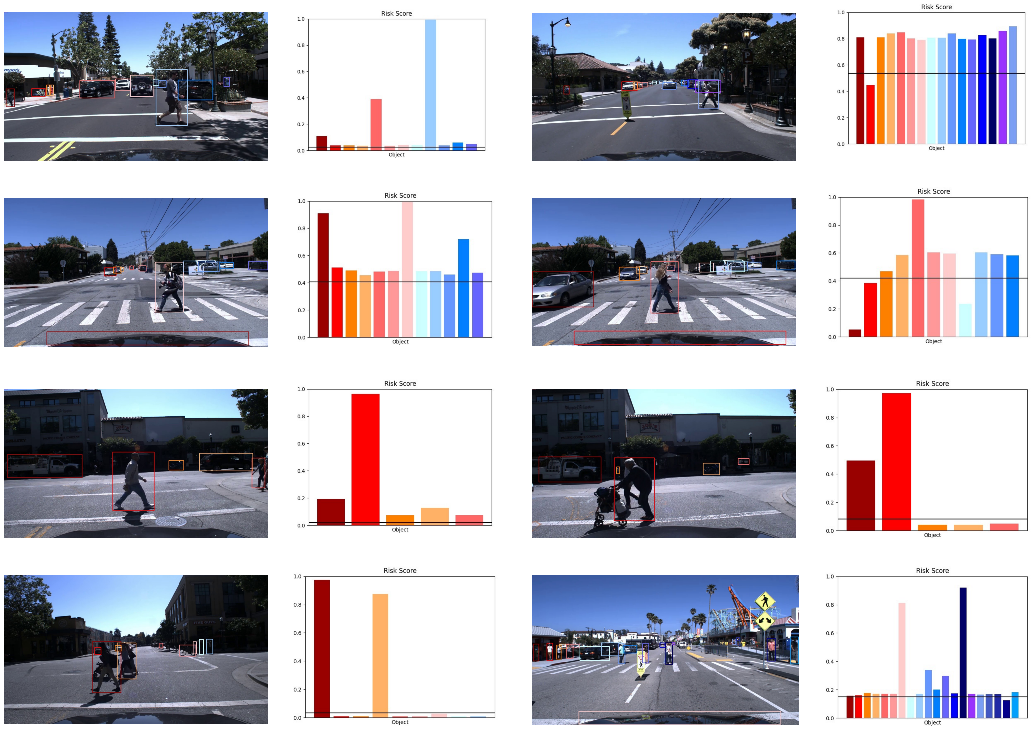


Figure 6: **Qualitative results of risk object identification – Crossing-Pedestrian.** We demonstrate the effectiveness of the proposed scene-level representation for risk object identification. According to the definition of the risky object proposed in [5], the candidate with the highest risk score is the risk object. In this figure, we show the risk scores of each object candidate and demonstrate the system can differentiate risk and non-risk objects in various crossing pedestrian scenarios. For each candidate, the color of the bar matches the color of the bounding box.

References

[1] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[2] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko. Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and Causal Reasoning. In *CVPR*, 2018.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *CVPR*, 2020.

[4] A. Jain, H. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models. In *ICCV*, 2015.

[5] C. Li, S. H. Chan, and Y.-T. Chen. Who Make Drivers Stop? Towards Driver-centric Risk assessment: Risk Object Identification via Causal Inference. In *IROS*, 2020.