Supplementary - Universal Embeddings for Spatio-Temporal Tagging of Self-Driving Logs

1 Model Architecture

Figure 1 provides the architecture details of our log embedding network. Our network is composed of the following components:

- `Conv(C)` represents a convolutional block that outputs `C` channels. It is a 2D convolutional layer followed by a ReLU activation (except for the final layer). For all convolutional layers, we use 3×3 convolutional kernels, with a stride of 1 and padding of 1, maintaining the original input resolution.
- `MaxPool` represents a `MaxPool` layer with a 3×3 kernel size, a stride of 2 and padding of 1. As a result, this layer returns a feature map with exactly half the spatial resolution.
- `AvgPool(K)` represents a `AvgPool` layer with `K` $\times$ `K` kernel size, with a stride of `K` and no padding.
- `Upsample 2x` represents upsampling feature maps to twice the resolution using bilinear interpolation.

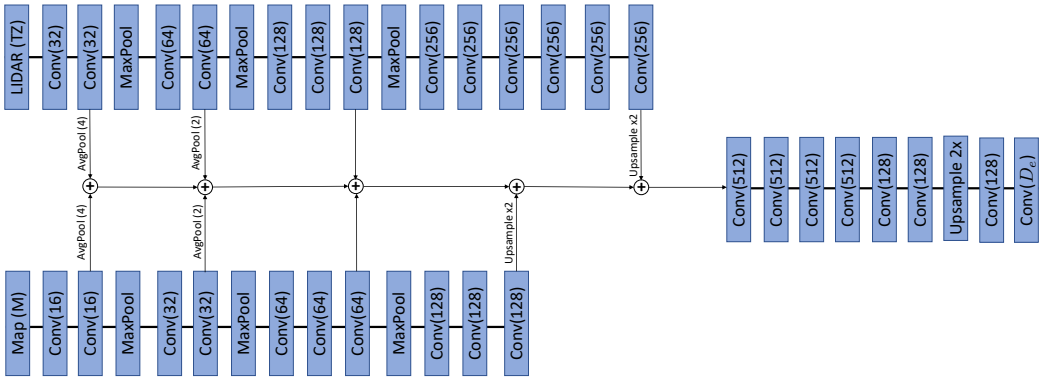

Figure 1: Model Architecture Details. "+" represents concatenation.

2 Interaction Definitions and Spatial Regions

In this section, we provide detailed definitions for our interaction attributes, "blocked by vehicle" and "braking on vehicle" used in our experiments. Note that while we use these particular definitions in our experiments, our method is general and can be trained to tag other vehicle interactions that are defined differently.

Vehicle Lane Association: For both interaction definitions, we require reasoning about whether any pair of vehicles are in the same lane. To determine this, we first compute which lanes each vehicle belongs to. We consider a vehicle *in a lane* if at least 20% of its bounding box overlaps with the lane's polygon. Given each vehicle's set of lanes, we consider two vehicles in the same lane if the intersection of their lane sets is not empty.

4th Conference on Robot Learning (CoRL 2020), Cambridge MA, USA.

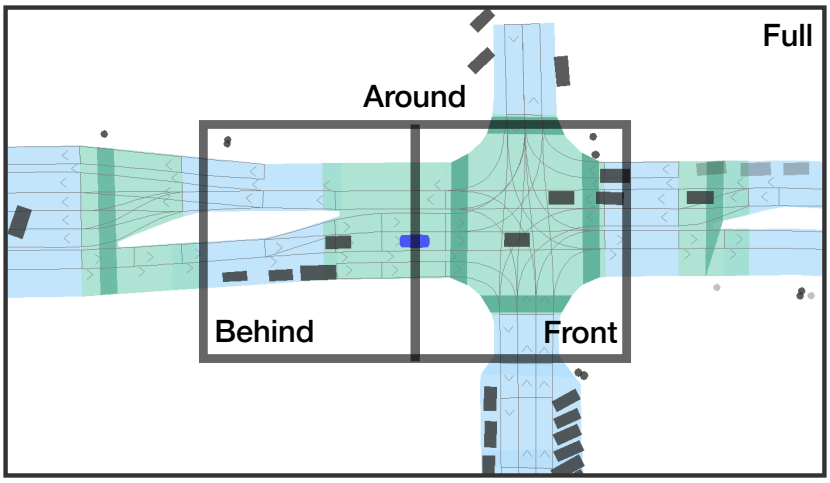

Figure 2: **Spatial Regions:** Visualization of the spatial regions used for evaluation.

Density		Actions										Map		Speed
Veh	Ped	P	S	B	KL	RT	LT	RC	LC	BB	BF	3-way	4-way	Speed
99	78	60	77	63	91	18	18	7	9	19	22	86	86	98

Table 1: **Dataset Statistics:** Percentage of frames in the training dataset with a positive example for each tag attribute.

Blocked By Vehicle: We consider a vehicle "blocked by a vehicle" if both of the following conditions are met:

- The target vehicle is stopped (i.e., vehicle speed ≤ 0.2 m/s)
- The target vehicle has another vehicle ahead *in its lane* and the bumper to bumper distance between the two vehicles is less than 5 meters.

Braking On Vehicle: We consider a vehicle "braking due to another vehicle" if both of the following conditions are met:

- The target vehicle is braking. (i.e., vehicle acceleration ≤ -0.6 m/s^2)
- The target vehicle has another vehicle ahead *in its lane* and the bumper to bumper distance between the two vehicles is less than 5 meters.

Spatial Regions: In Table 1 of the main paper, we evaluate our model's ability to produce scene tags for 4 SDV-relative regions, R. We visualize these regions in Figure 2. The four regions are defined as follows:

- *Full*: The entire scene, which we define to be the 140 meter by 80 meter rectangle centered at the SDV.
- *Around*: The 70 meter by 40 meter rectangular region centered at the SDV.
- *Front*: All areas from the *Around* region that are in front of the SDV.
- *Behind*: All areas from the *Around* region that are behind the SDV.

3 Dataset Statistics

Positive Examples: To showcase the importance of our task balancing scheme, we demonstrate the imbalances between the supervision available for each task by computing dataset statistics. In Table 1, we compute the percentage of timesteps with at least one positive example for each tag.

		Full		Around		Front		Behind		Intersection		Crosswalk
Pixel	Global	Veh	Ped	Veh	Ped	Veh	Ped	Veh	Ped	Veh	Ped	Ped
	✓	2.34	2.35	3.21	1.88	1.90	1.34	1.77	0.95	3.19	1.09	0.86
✓		0.97	1.43	0.48	0.92	0.32	0.62	0.31	0.46	0.32	**0.27**	**0.21**
✓	✓	**0.93**	**1.40**	**0.46**	**0.90**	**0.31**	**0.61**	**0.29**	**0.45**	**0.31**	**0.27**	**0.21**

Table 2: **Density Losses:** L1 error for density tags across all spatial locations for models trained with different density losses.

Hyperparameters		Performance	
K	α	Discrete (F1) ↑	Continuous (L1) ↓
3	10^{-4}	72	0.93
2	10^{-4}	72	0.98
4	10^{-4}	72	0.93
3	10^{-3}	73	1.06
3	3×10^{-3}	73	0.88
3	3×10^{-4}	69	1.05

Table 3: **Hyperparameter Sensitivity:** Final performance of models retrained with different hyperparameter settings for the hard negative mining ratio, K, and the initial learning rate, α.

For continuous attributes, we consider any scene that has a spatial location with non-zero values as a positive example.

4 Additional Ablations

Global Loss Ablation: As mentioned in the main paper, to train density attributes, we use an additional "global" loss that is applied after pooling via summation,

$$\ell_\tau(\mathbf{V}^{(t)}, \tilde{\mathbf{V}}^{(t)}) = \left| \sum_{h,w} \mathbf{V}^{(t)}[h,w] - \sum_{h,w} \tilde{\mathbf{V}}^{(t)}[h,w] \right| \quad . \tag{1}$$

This helps reduce spatially correlated errors which can have negative effects when summed over larger regions. For example, if each spatial location underestimates the density, these errors will be accumulated during pooling. Table 2 shows better density performance with this additional loss.

Hyperparameter Sensitivity: In Table 3, we retrained our model with different settings for the learning rate α and hard negative mining ratio, K. Generally, we find our results are not very sensitive to these hyperparameter choices.

5 Additional Qualitative Results

Sample Answers: Figure 3 shows additional scene tags generated by our model, in the same format as the main paper. We include tags relating to density, actions, interactions, vehicle speed, and map topology. We use a variety of spatial regions for the tags (each defined above). The results demonstrate that our model can handle this diverse set of tag attributes. We include both positive results and failure cases.

Retrieval: Figures 4-8 show additional qualitative examples of our model in the retrieval setting. Each figure shows visualizations of the scenes with actor labels shown (not given to the model, but useful for qualitative evaluation) that have scene tags that match some criteria. For example, Figure 4 shows scenes with a tag value greater than 5 for the pedestrian density in front of the SDV. Our results demonstrate that our model can retrieve useful scenes based on a variety of conditions relating to pedestrian density, vehicle density, intersection types, and vehicle actions.

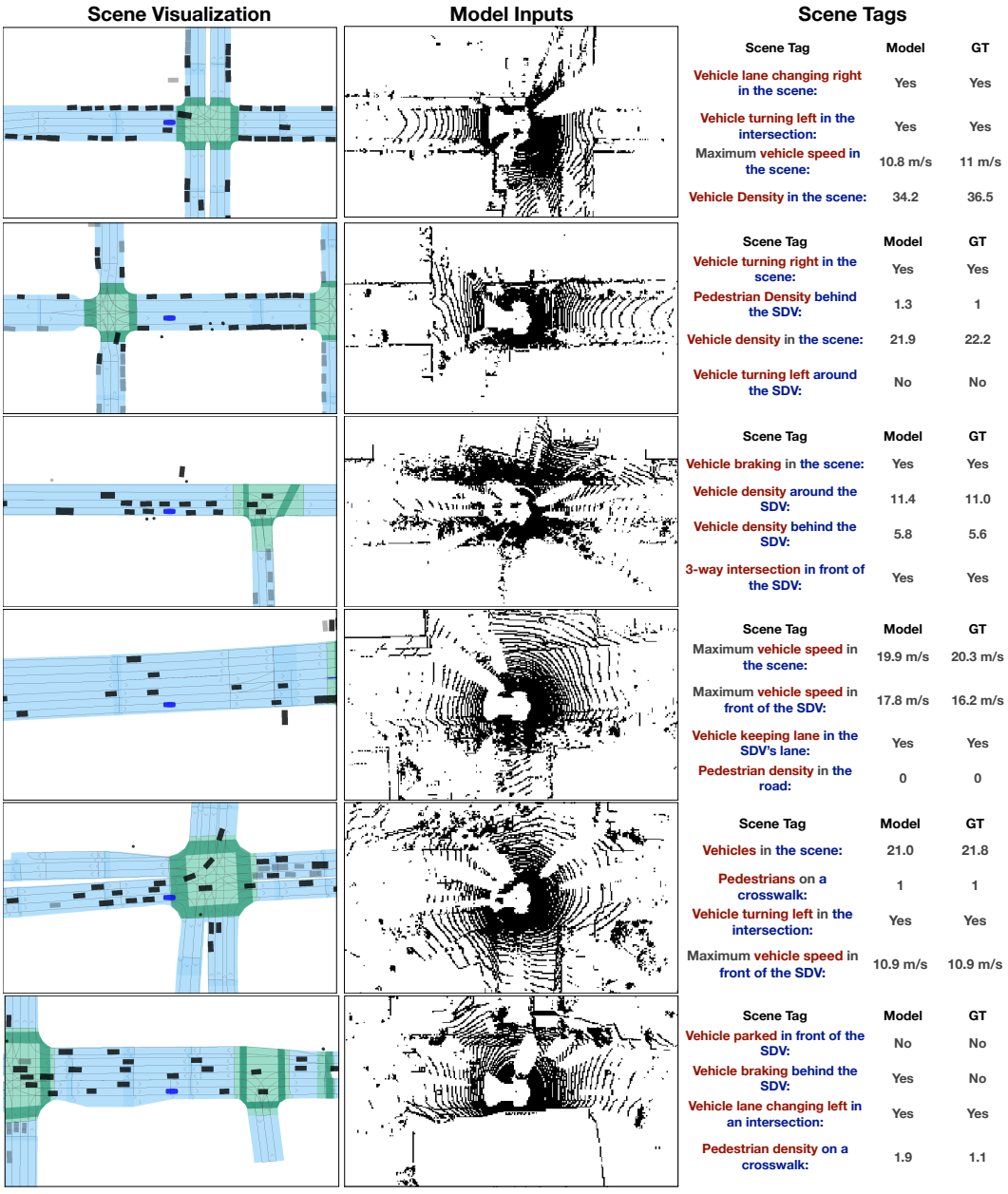

Figure 3: **Additional scene tags:** (Left) We display both the rasterized map (input to the model) and actor labels (**not** used by the model). The SDV is dark blue, visible actors are black, occluded actors are grey, and pedestrians are circles. (Center) Sensor observations are shown for the central timestep. (Right) Scene tags generated by our model. Tag attributes, τ, are in red, and regions, R are in blue. Tags shown across multiple attributes and regions of interest.

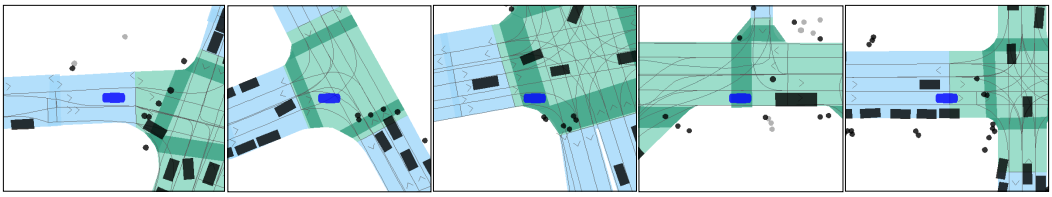

Figure 4: **Retrieval (Pedestrians):** Visualizations of scenes retrieved by our model from the evaluation set with 5 pedestrians or more in front of the SDV.

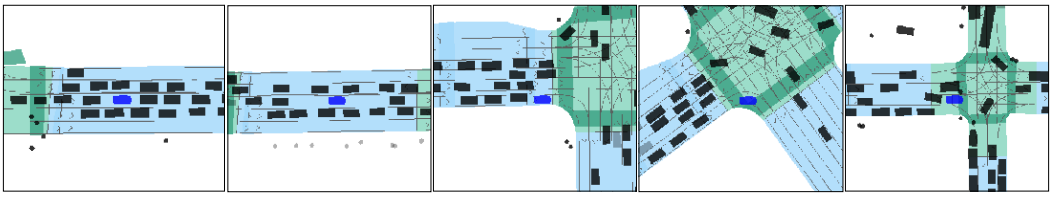

Figure 5: **Retrieval (Vehicles):** Visualizations of scenes retrieved by our model from the evaluation set with 10 vehicles or more around the SDV.

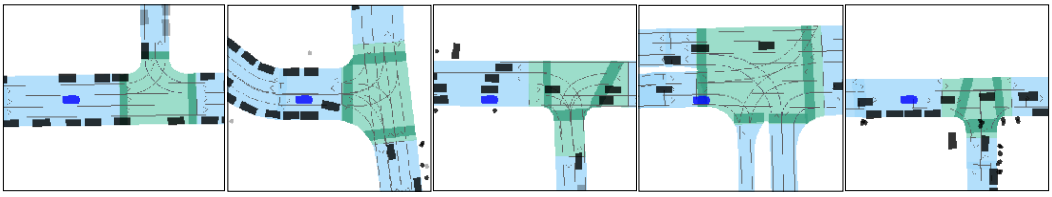

Figure 6: **Retrieval (3-Way):** Visualizations of scenes retrieved by our model from the evaluation set with 3-way intersections in front of the SDV.

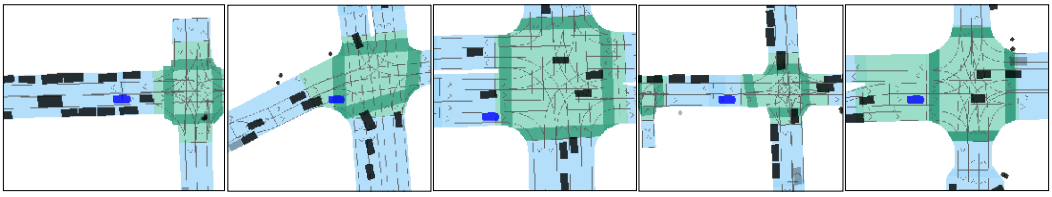

Figure 7: **Retrieval (4-way):** Visualizations of scenes retrieved by our model from the evaluation set with 4-way intersections in front of the SDV.

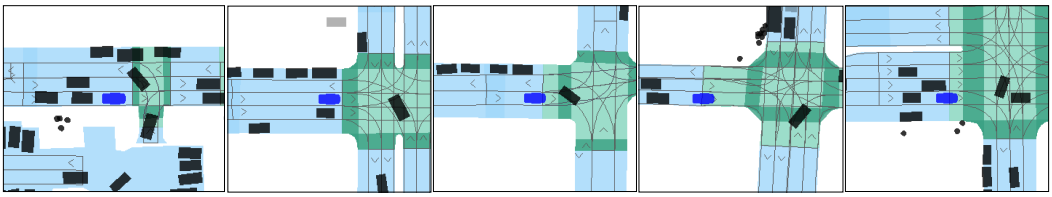

Figure 8: **Retrieval (Turns):** Visualizations of scenes retrieved by our model from the evaluation set with vehicles turning in front of the SDV.