Supplementary Material

TABLE I
COMPARISON OF TWO METHODS OF OBJECT REPRESENTATION. 'SF' FEEDS THE SEMANTIC FEATURES OF BACKBONES INTO THE LSTM, WHILE 'HF' USING HIGH-DIMENSIONAL FEATURES. BEST RESULTS IN BOLD.

	Car	Pedestrian	Motorcycle	Overall Accuracy
SF	96.06	86.86	88.72	91.48
HF	**98.59**	**100.0**	**97.44**	**98.73**

I. Overview

In this supplementary material, we first conduct the experiment of embedding high-dimensional features or categories in Section II-A. Besides, we analyze the characteristics of LSTM and MLP in Section II-B.

II. Additional Experiment

A. Experiment on Methods of Object Representation

As for the strategy of representing object features, we have two methods, the category-based and the high-dimensional feature-based. The category-based one encodes point clouds by concatenating the semantic outputs of several networks. To provide more detailed information for subsequent classification, the outputs of each network retain the top 3 categories with the highest scores, forming a 9-dimensional vector in total. This vector has more detailed and intuitive information than the original point clouds. The high-dimensional feature-based one exploits 256-dimensional features of the fully connected layer in each backbone. After the feature extraction, we adopt the feature-level fusion to integrate the object information. The specific method is to form a feature vector which combines the 768-dimensional feature.

As shown in Table I, the approach of using high-dimensional features demonstrates satisfying classification results than using semantic features. In this way, we obtain the +13% accuracy improvement of pedestrians and +9% accuracy improvement of motorcycles. Although the semantic features are detailed and intuitive, it is hard to classify motorcyclist and pedestrian in some cases due to their highly similar semantic outputs. Instead, high-dimensional features retain higher representation ability and more information which may enhance the difference between pedestrians and motorcycles.

B. Experiment on Time-series Fusion

Fusion Model. In time-series fusion classification, we conduct experiments based on multi-layer perceptron and long short-term memory. When resorting to multi-layer perceptron, by adjusting the number of hidden layers and the number of neurons in each hidden layer, we find out that the multi-layer perceptron with three hidden layers of 2048, 1024, and 256 obtains the best classification result. According to Table II the overall classification accuracy achieves 96.95%. In long short-term memory experiment, we design three forms of 1.5 seconds, namely 15, 5, and 3 input moments and the former one is verified to be the optimum. In the network structure, we exploit the hidden vector size of 512 and the embedding dimension of distances is selected as 128, 256, or 512. Furthermore, we improve the network generalization by setting up Dropout of 0.2. The experiment demonstrates that embedding the distance to 128 dimensions obtains a best overall classification accuracy of 98.73%. It well verifies that LSTM shows its advantage of gradual classification.

TABLE II
COMPARISON OF TIME-SERIES FUSION RESULTS. MLP AND LSTM ARE SELECTED AS THE NETWORK STRUCTURES. SOFTMAX, BINARY CROSS ENTROPY LOSS AND FOCAL LOSS ARE TESTED. BEST RESULTS IN BOLD.

	Car	Pedestrian	Motorcycle	Overall Accuracy
MLP+Softmax-CE	99.15	90.25	**97.95**	96.18
MLP+Binary-CE	98.59	100.0	90.26	96.95
MLP+Focal-Loss	98.31	100.0	84.62	95.42
LSTM+Softmax-CE	**99.44**	89.83	97.44	96.06
LSTM+Binary-CE	98.59	99.58	90.26	96.82
LSTM+Focal-Loss	98.59	**100.0**	97.44	**98.73**

TABLE III
COMPARISON BETWEEN DIFFERENT VARIANTS OF OUR APPROACH WITH / WITHOUT THE DISTANCE BETWEEN OBJECTS AND THE SENSOR. ONE METHOD IS TO SET A DISTANCE THRESHOLD IN THE OUTPUT OF LSTM. ANOTHER METHOD ADDS DISTANCE TO THE INPUT OF LSTM. BEST RESULTS IN BOLD. OA REFERS TO THE OVERALL ACCURACY.

	Insert Location	Car	Pedestrian	Motorcycle	OA
MLP	/	98.31	99.15	81.54	94.40
LSTM	/	98.03	99.15	85.64	95.29
MLP + dist(10)	Output	98.31	98.73	85.64	95.17
MLP + dist(12)	Output	98.31	99.15	82.05	94.53
LSTM + dist(10)	Output	98.31	99.58	90.26	96.69
LSTM + dist(12)	Output	98.59	99.58	87.18	96.06
MLP + dist	Input	98.31	100.0	84.62	95.42
LSTM + dist	Input	**98.59**	**100.0**	**97.44**	**98.73**

It is believed that multi-layer perceptron essentially processes the entire object information in a certain period of time and finds the best classification standard through the combination of information and the distribution of weights. However, compared with the former, long short-term memory has the ability of gradual classification. Different from the multi-layer perceptron focusing on the global information, it pays more attention to the relationship of adjacent inputs. The high-dimensional features corresponding to the adjacent scans of dynamic objects change during movement, so as to learn the change rules. For example, the first two moments of the classification results show low confidence and the overall range is below 0.3. With the participation of subsequent inputs, the confidence increases obviously, rising from 0.5 to 0.7, eventually maintain above 0.9. It well verifies that LSTM shows the advantages of its gradual classification when classifying dynamic objects.

Distance of Dynamic Objects. As shown in Table III, We also consider the effect of adding the distance between objects and the LiDAR on the classification accuracy. Firstly, we set the distance threshold in the output of the LSTM so as to find the optimal distance boundary. If the distance exceeds this threshold, we only divide the object into pedestrian and car In the specific implementation, in addition to modifying the classification accuracy according to the above criteria, we also need to ignore the loss of the motorcycle when the distance exceeds the threshold. In this paper, the distance threshold is set to be 10/12 meters and the results show that 10 meters is the optimal distance threshold. Differently, the 'input' means that the distances are fed into LSTM along with three high-dimensional features as the input, improving the classification accuracy of motorcyclists to 98.73%. Considering that LSTM is a gradual classification network, it can learn the speed difference of objects according to the change of distances.