Simultaneous Estimation of Driving Intentions for Multiple Vehicles Using Video Transformer

Junya Isogawa, Fumihiko Sakaue and Jun Sato
Nagoya Institute of Technology, Nagoya 466-8555, Japan
{*isogawa@cv., sakaue@, junsato@*}*nitech.ac.jp*

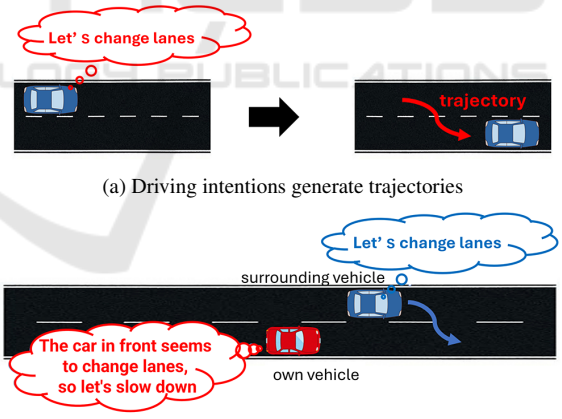
Keywords: Driving Intention, Multiple Vehicles, Multiple Drivers, Vehicle Trajectory, Video Transformer.

Abstract: In autonomous driving, it is important for the vehicle to appropriately determine the next action to be taken on the road. In complex situations such as on public roads, the better action for the own vehicle can be determined by considering the driving intention of other vehicles around the vehicle. Thus, in this paper, we propose a method to determine the next action of the own vehicle by simultaneously estimating the next driving intentions of all vehicles, including other vehicles around the own vehicle. The time series of vehicle motions on the road can be represented as sequential images centered on the vehicle. In this paper, we analyze the sequential images of vehicle trajectories using the Video Transformer and simultaneously predict the driving intentions of all vehicles on the road. In general, driving intentions change over time. Thus, in this research, we first propose a method to predict the next intention, and then extend it to predict the transition of driving intentions over the next few seconds. We also apply our method to predict driving trajectories, and show that the prediction of the driving trajectory can be improved by using the driving intentions estimated from the proposed method.

1 INTRODUCTION

Autonomous driving technology is very important for reducing traffic accidents and improving the efficiency of logistics, and many research and development projects (Bojarski et al., 2016; Codevilla et al., 2018; Grigorescu et al., 2020) have been conducted on it. Autonomous driving is realized by bringing together technologies from various fields, and among them, determining the optimal behavior from observing traffic conditions is an essential technology for autonomous driving.

In the existing research on vehicle behavior decisions, people have determined the vehicle's trajectory from the observed traffic conditions (Jeong et al., 2017; Hegde et al., 2020; Richardos et al., 2020). However, when a human drives a vehicle, as shown in Figure 1 (a), first the driving intention, such as "let's change lanes," is determined, and then a driving trajectory is generated based on the intention. Therefore, in this research, we consider the driving intention to be important information that is the source of trajectory generation, and propose a method to determine the driving intention that the vehicle should take next. By accurately determining the next intention to be taken in this way, the trajectory generation in autonomous driving can be made more accurate.



(a) Driving intentions generate trajectories

(b) Interaction of intentions among vehicles

Figure 1: (a) Driving intention is important information that is the basis for trajectory generation. (b) By predicting the driving intentions of surrounding vehicles, we can determine better driving intentions for the own vehicle.

The intention of the vehicle is not determined independently by the vehicle. For example, in Figure 1 (b), the vehicle ahead is trying to change lanes, so the vehicle generates an intention to decelerate. In this way, the vehicle's intention is determined taking into

Isogawa, J., Sakaue, F. and Sato, J.
Simultaneous Estimation of Driving Intentions for Multiple Vehicles Using Video Transformer.
DOI: 10.5220/0013232100003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 471-477
ISBN: 978-989-758-728-3; ISSN: 2184-4321

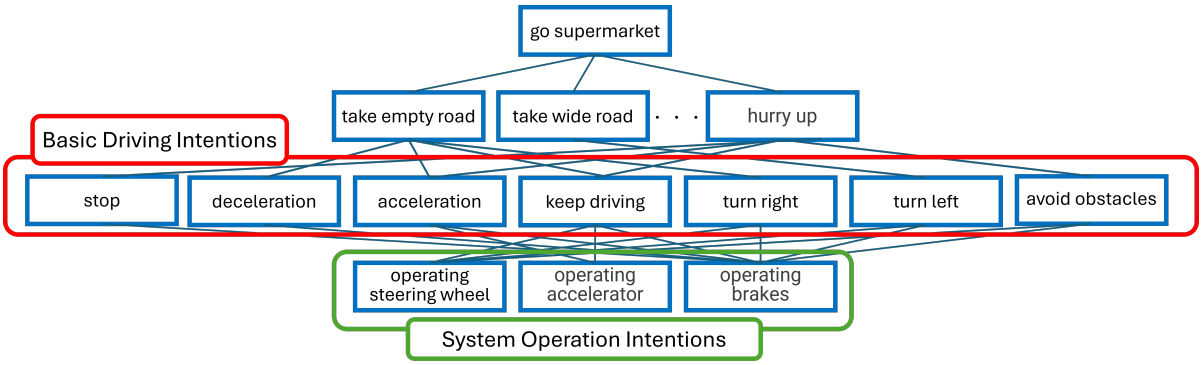


Figure 2: Hierarchy of driving intentions. Driving intentions range from high-level intentions to low-level intentions, and these are thought to have a hierarchical structure. The intentions that are directly required for the autonomous driving system are the system operation intentions. In order to determine the system operation intentions, it is important to appropriately determine the basic driving intentions shown in the red frame. In this research, we estimate these 7 basic intentions from vehicle trajectories. Complex behavior in various situations on public roads can be expressed by combinations of these 7 basic driving intentions.

account the intentions of other vehicles. Therefore, each vehicle can determine better driving behavior by predicting the driving intentions that the surrounding vehicles will take at the next time.

Research on predicting trajectories that consider interactions among multiple vehicles has been conducted in the past (Mo et al., 2020; Zhang et al., 2020; Dai et al., 2021; Chen and Krahenbuhl, 2022). However, these studies aimed to estimate vehicle trajectories, and did not estimate the driver's intention, which is the important source of these vehicle trajectories.

Thus, in this research, we propose a method to simultaneously predict the intentions of the vehicle and surrounding vehicles while considering interactions between vehicles. By extracting the driving intention of the own vehicle from the predicted intentions of all vehicles, the own vehicle can determine the best driving intention to be taken next even in complex situations.

Technologies for estimating driving intentions have been proposed in the past (Han et al., 2019; Huang et al., 2022). However, these were limited to estimating simple intentions such as lane changes on straight roads and right and left turns at intersections, and could not handle driving intentions in complex situations such as waiting for a straight vehicle coming from ahead before turning right at an intersection. In this research, we estimate the best driving intention that the own vehicle should take in such a complex situation by simultaneously estimating the driving intentions of the surrounding vehicles along with the driving intention of the own vehicle. By using the driving intentions estimated in this way, it is expected that we can generate more accurate driving trajectories.

2 SIMULTANEOUS PREDICTION OF DRIVING INTENTIONS FOR MULTIPLE VEHICLES

2.1 Driving Intentions

We first discuss the driving intentions we consider in this paper.

In general, there are various levels of intentions that drivers have while driving (Azad et al., 2020; Albrecht and Bartneck, 2019). These range from high-level intentions such as "I want to go to the supermarket" to low-level intentions such as "Turn the steering wheel to the right", and these are thought to have a hierarchical structure as shown in Figure 2.

In autonomous driving, the intentions that are directly required for the system to operate the steering wheel and brakes are the system operation intentions shown in the green frame in Figure 2. In order to determine these system operation intentions, it is important to appropriately determine the driving intentions shown in the red frame that are located just above the system operation intentions and are the basis for deciding these intentions. Therefore, in this research, we define 7 basic driving intentions, which are "stop", "deceleration", "acceleration", "keep driving", "turn right", "turn left", and "avoid obstacles", and consider estimating them. Complex behavior in various situations on public roads can be expressed by combinations of these 7 basic driving intentions. For example, when making a right turn at an intersection and an oncoming vehicle is approaching, the vehicle will first generate the intention to "stop" at the intersection, generate the intention to "turn right"

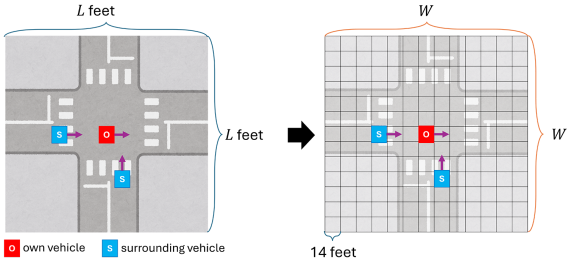


Figure 3: Network input.

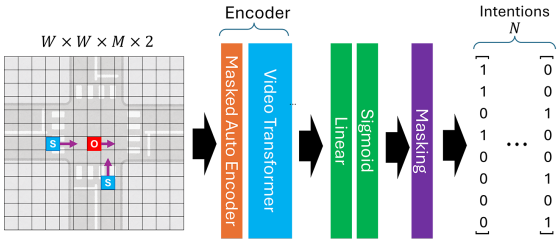


Figure 4: Network for intention prediction.

once the oncoming vehicle has passed, and generate the intention to "keep driving" once the right turn is complete. In this research, we estimate the combination of these 7 basic intentions that the vehicle should take and their transition over time. In the following, the number of basic intentions will be referred to as I. That is $I = 7$.

2.2 Network Input and Output

Next, we explain the input to the network. In this research, in order to handle various road conditions on general roads, such as straight roads and intersections, we consider a square area of L feet $\times$ L feet centered on the own vehicle at each time instant, as shown in Figure 3, and input the trajectories of all vehicles within this area. In this research, vehicles other than the own vehicle within this area are called surrounding vehicles. This area is discretized into 14 feet, which corresponds to the size of one vehicle, and expressed as a $W \times W$ array. In this research, we set $L = 182$ feet, and $W = 13$, so the total number N of array elements is 169 $(= 13 \times 13)$.

Then, the trajectory of each vehicle in T_I seconds is divided into M, and M pairs of X and Y coordinates of the trajectory is stored in the vehicle position in the array. This array is considered to be a time-series image tensor of $W \times W \times M \times 2$, and is used as the input to our network. In this research, we consider vehicle trajectory in $T_I = 6$ seconds and it is divided into $M = 60$ for every 1/10 second.

The output is the intentions occurring in the T_O seconds following the input trajectory for each vehicle in the $W \times W$ image area. In other words, it is the probability for each of the seven driving intentions defined in the previous section. If the probability of a certain intention for a certain vehicle exceeds a threshold, the vehicle is considered to have the intention, and a multi-hot vector of intentions is obtained. In this way, a set of driving intentions the vehicle should take in the next T_O seconds is determined at each pixel of the $W \times W$ image area. In this research, we set $T_O = 5$ seconds.

2.3 Network Structure and Training

Next, we describe the network structure and training for predicting driving intentions. The overall network structure is shown in Figure 4. The network first inputs a $W \times W \times M \times 2$ image tensor into an encoder consisting of Masked Auto Encoder (He et al., 2021) and Video Transformer (Bertasius et al., 2021) to obtain the features of the vehicle trajectory. Then, the probability for each driving intention is computed using a fully connected layer and a sigmoid function prepared for each driving intention. Finally, the obtained probabilities are thresholded to obtain a multi-hot vector of the driving intention at each pixel of the N $(= W \times W)$ image area.

This network is trained by minimizing the loss $Loss1$, which consists of the cross entropy between the true value and the estimated value of all intentions for all image pixels, as follows:

$$Loss1 = -\sum_{n=1}^{N}\sum_{i=1}^{I} p_v(n,i)\log q_v(n,i) \quad (1)$$

where, $p_v(n,i)$ and $q_v(n,i)$ represent the true and estimated probability of the driving intention i at the nth pixel, respectively.

2.4 Predicting Temporal Transition of Intention

So far, we have explained a method that outputs a multi-hot vector of driving intentions, but this method derives only a combination of basic intentions to be taken in a target scene, and it is unclear in what order and at what timing these should be taken. Therefore, next we will describe a method to derive the temporal transition of multiple intentions to be taken in each situation.

For deriving the temporal transition of multiple intentions, no thresholding is performed, and the intent probability is output at each time instant, as shown in Figure 5. In this research, each T_O-second data is divided into M_O, and the transitions of driving intention over T_O seconds are predicted every T_O/M_O second. In this research, we set $T_O = 5$ seconds and $M_O = 50$.

Table 1: Definition of intention.

intention	definition
stop	speed is less than 10km/h
deceleration	speed decreases by more than 10km/h per second
acceleration	speed increases by more than 10km/h per second
turn right	direction changes to the right by more than 6 degrees per second
turn left	direction changes to the left by more than 6 degrees per second
avoid obstacles	vehicle in front moves behind the own vehicle
keep driving	speed is more than 10km/h and speed change is less than 10km/h

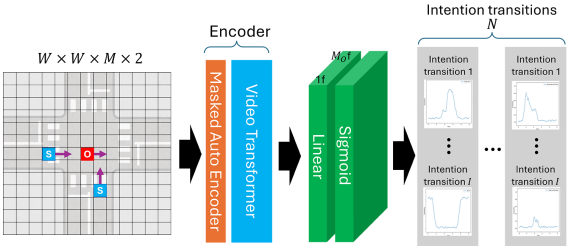


Figure 5: Network for intention transition prediction.

The network is trained by minimizing the loss $Loss2$, which is the sum of the cross entropy errors of intentions for all pixels at all times, as follows:

$$Loss2 = -\sum_{m=1}^{M_O}\sum_{n=1}^{N}\sum_{i=1}^{I} p_v(m,n,i)\log q_v(m,n,i) \quad (2)$$

where, $p_v(m,n)$ and $q_v(m,n)$ represent the true and estimated probability distributions of the driving intention at the nth pixel at time m, respectively.

This method can estimate the transition of driving intentions every T_O/M_O second, making it possible to determine the transition of system operation intentions. Since the proposed method simultaneously predicts the driving intentions of all vehicles, it enables us to predict the transition of driving intention in complex situations where we need interaction with other vehicles.

3 PREDICTING DRIVING TRAJECTORY BASED ON DRIVING INTENTION PREDICTION

The driving intentions predicted by the proposed method can be used not only as the next driving behavior to be taken by the vehicle, but also to predict the driving trajectory of the vehicle. In this section, we show a method to improve the accuracy of estimating the driving trajectory of vehicles by using the driving intentions of vehicles estimated using the proposed method.

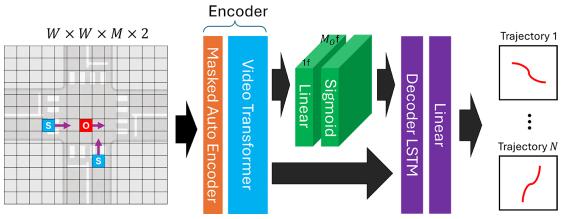


Figure 6: Network for trajectory prediction.

In this method, as shown in Figure 6, an LSTM Decorder (Hochreiter and Schmidhuber, 1997) and a fully connected layer are added to the network that predicts the driving intentions, and vehicle trajectories are predicted from the predicted driving intentions obtained by the method described in the previous section and the features extracted by the Encoder. The predicted trajectory (time-series position information) of each vehicle is computed by dividing T_O seconds of data into M_O, and saving pairs of X and Y coordinates every T_O/M_O seconds. Again, we set $T_O = 5$ seconds and $M_O = 50$.

To train this network, the trajectory loss $Loss\_traj$ is defined as follows:

$$Loss\_traj = \sqrt{\frac{1}{N \times M_O}\sum_{n=1}^{N}\sum_{m=1}^{M_O} ||\mathbf{x}_{nm} - \hat{\mathbf{x}}_{nm}||^2} \quad (3)$$

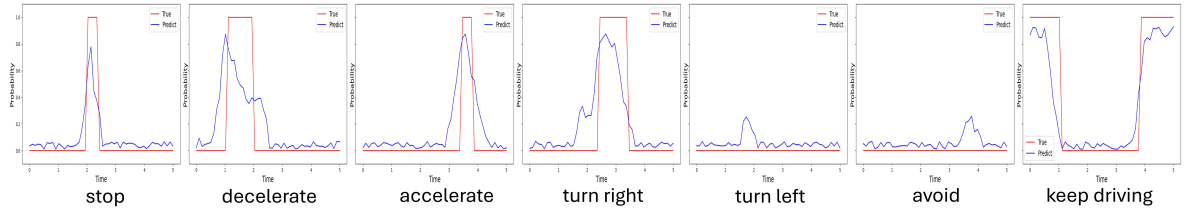
where, $\mathbf{x}_{nm}$ is the true values of the X and Y coordinates of vehicle n at time m, and $\hat{\mathbf{x}}_{nm}$ is the coordinates predicted by the network. By adding this trajectory loss $Loss\_traj$ to the intention loss $Loss2$ defined in equation (2), we define the loss $Loss3$ for training the trajectory prediction network as follows:

$$Loss3 = Loss2 + Loss\_traj \quad (4)$$

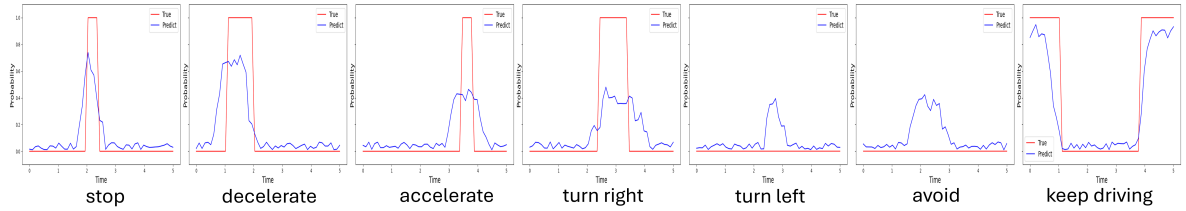
The trajectory prediction network is trained by minimizing $Loss3$.

4 DATASET

To train the proposed intention prediction network and trajectory prediction network, we need a large



(a) proposed.



(b) Own vehicle only.

Figure 7: Predicted transitions of intention.

Table 2: Accuracy of driving intention prediction.

	stop	decel.	accel.	driving	turn right	turn left	avoid	total
proposed	85.31	80.26	81.01	89.44	79.13	73.81	76.44	85.08
own vehicle only	86.27	77.43	78.11	89.49	77.06	70.22	78.77	82.61

number of pairs of trajectories and intentions of the own vehicle and surrounding vehicles. In this research, we use Argoverse2 (Wilson et al., 2023) as the basis dataset, which consists of real vehicle trajectories captured at 10 Hz for 11 seconds. In this research, we used the first 6 seconds in the 11-second data to estimate the driving intentions and trajectories for the following 5 seconds. Since there is no intention information in this dataset, we added the intention of each vehicle based on the trajectory of the vehicle.

Intent information is added in a rule-based manner based on the trajectory information. For each of the driving intentions, the definitions shown in Table 1 are applied, and intention information is generated from the trajectory information following these definitions. Since each trajectory data may contain multiple intentions, we allowed multiple driving intentions to be set for each data.

In this way, a dataset was constructed by creating 12,000 pairs of data consisting of trajectories and intentions for the own vehicle and surrounding vehicles.

5 EXPERIMENTS

Using the proposed method, we predicted the intentions of all vehicles based on the past trajectories of the own vehicle and its surrounding vehicles.

10,000 data were randomly extracted from the created dataset to serve as the training dataset, and the remaining 2,000 data were used as test data. The batch size for training the network was 128, the learning rate was 10^{-6}, and training was performed for 100 epochs. Adam (Adaptive moment estimation) (Kingma and Ba, 2014) was used to optimize the training. Using the network trained in this way, we predicted the intentions of the test data and compared them with the true intentions of the test data. For comparison, we also constructed a model that only predicts the intentions of the own vehicle, and compared its results with those of the proposed method.

5.1 Prediction of Driving Intention

In Table 2, we show the prediction accuracy when predicting multiple intentions contained in each 5-second data as a multi-hot vector. The table shows that the proposed method, which simultaneously predicts the intentions of all vehicles, has significantly higher prediction accuracy for each of the 7 intentions than the method that estimates only the own vehicle intentions. The proposed method also has a significantly higher accuracy in the total score of all seven intentions. These results confirm the effectiveness of the proposed method, which simultaneously predicts vehicle intentions for estimating driving intentions.

Next, we show the results of predicting the transition of intention probability in Figure 7. Figure 7 (a) shows the transition of intentions estimated by the proposed method, and Figure 7 (b) shows the result of

Table 3: RMSE of predicted intention transitions.

	RMSE
proposed	0.313
own vehicle only	0.322

Table 4: RMSE of predicted trajectories.

	RMSE (feet)
proposed	4.92
own vehicle only	5.01

predicting only the own vehicle. The red line in the figure is the true value of the transition of intentions, and the blue line is the estimated value. From these figures, we find that, for almost all intentions, the proposed method can estimate the transition of intentions closer to the true value.

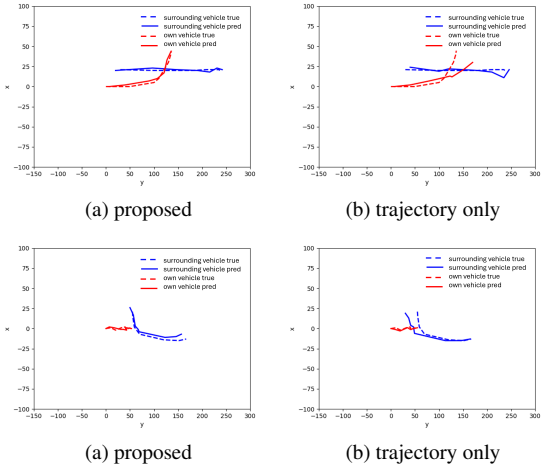
Table 3 shows the results of computing the RMSE of the predicted intention transition against the true intention transition for all test data. As shown in this table, the effectiveness of the proposed method for simultaneously predicting vehicle intentions can also be confirmed numerically.

5.2 Prediction of Driving Trajectory

Next, we show the experimental results of the trajectory prediction using the intention predicted by the proposed method. For comparison, we also show the results of a network that predicts only the trajectory without using the intention prediction.

Figure 8 (a) shows the prediction results of the proposed method, and Figure 8 (b) shows the prediction results when predicting only the trajectory. The red line represents the trajectory of the own vehicle, and the blue line represents the trajectory of the surrounding vehicles. The solid lines represent the true trajectories of each vehicle, and the dashed lines represent the predicted trajectories. In the upper case of Figure 8, when predicting only the trajectory, the trajectory of the own vehicle with the intention of turning left is not predicted well, whereas the proposed method predicts the own trajectory correctly. In the lower case of Figure 8, when predicting only the trajectory, the trajectory of the other vehicle with the intention of turning left and merging is not predicted well, whereas the proposed method predicts it correctly.

Table 4 shows the RMSE of the predicted trajectories for all test data. In this table, the proposed method also shows improvements in accuracy compared to the method that predicts only the trajectory, which confirms that intention prediction using the proposed method is effective in improving the accuracy of trajectory prediction.



(a) proposed (b) trajectory only

(a) proposed (b) trajectory only

Figure 8: Prediction of driving trajectory.

6 CONCLUSION

In this paper, we proposed a method to determine the next driving intention of the own vehicle by simultaneously predicting the driving intentions of all vehicles. We also applied the proposed method for predicting intentions to improve the accuracy of vehicle trajectory prediction.

In the proposed method, we created a new training dataset containing true intentions and used this for training and testing of our network, demonstrating that it is possible to predict the intention of the own vehicle with high accuracy while taking into account interactions among multiple vehicles. We also applied the proposed method for predicting intentions to trajectory prediction, and showed that the accuracy of trajectory prediction can be improved by using intention prediction.

Although our method is still in its early stages, our method to simultaneously predict the driving intentions of all vehicles is promising.

REFERENCES

Albrecht, A. M. R. and Bartneck, J. A. M. C. (2019). Intent recognition in human-robot interaction: A review. *International Journal of Social Robotics*, 11(3):355–369.

Azad, A. K., Tiwari, N., and et al. (2020). A survey on intention recognition in autonomous vehicles. *Journal of Autonomous Vehicles and Systems*, 1:1–15.

Bertasius, G., Wang, H., and Torresani, L. (2021). Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*. Accepted to ICML 2021.

Bojarski, M., Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba, K. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Chen, D. and Krahenbuhl, P. (2022). Learning from all vehicles. *IEEE Conference on Computer Vision and Pattern Recognition*.

Codevilla, F., Müller, M., López, A., Koltun, V., and Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4693–4700.

Dai, S., Li, L., and Li, Z. (2021). Modeling vehicle interactions via modified lstm models for trajectory prediction. *IEEE Access*, 7:2169–3536.

Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386.

Han, T., Jung, J., and Ozguner, U. (2019). Driving intention recognition and lane change prediction on the highway. In *IEEE Intelligent Vehicles Symposium*.

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., and Girshick, R. (2021). Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*. Tech report. v3: add robustness evaluation.

Hegde, C., Dash, S., and Agarwal, P. (2020). Vehicle trajectory prediction using gan. *Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. In *Neural Computation*, volume 9, pages 1735–1780.

Huang, H., Zeng, Z., Yao, D., Pei, X., and Zhang, Y. (2022). Spatial-temporal convLSTM for vehicle driving intention prediction. *Tsinghua Science and Technology*, 27:599–609.

Jeong, D., Baek, M., and Lee, S.-S. (2017). Long-term prediction of vehicle trajectory based on a deep neural network. *International Conference on Information and Communication Technology Convergence (ICTC)*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mo, X., Xing, Y., and Lv, C. (2020). Interaction-aware trajectory prediction of connected vehicles using cnn-lstm networks. *IECON The 46th Annual Conference of the IEEE Industrial Electronics Society*.

Richardos, D., Anastasia, B., Georgios, D., and Angelos, A. (2020). Vehicle maneuver-based long-term trajectory prediction at intersection crossings. *IEEE 3rd Connected and Automated Vehicles Symposium (CAVS)*.

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J. K., Ramanan, D., Carr, P., and Hays, J. (2023). Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks.

Zhang, T., Fu, M., Song, W., Yang, Y., and Wang, M. (2020). Trajectory prediction based on constraints of vehicle kinematics and social interaction. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*.