

Elsevier required licence: © <2020>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>
The definitive publisher version is available online at <http://doi.org/10.1016/j.robot.2020.103477>

DeepGoal: Learning to Drive with driving intention from Human Control Demonstration

Huifang Ma¹, Yue Wang^{a,1,*}, Rong Xiong^{a,1,*}, Sarath Kodagoda², Li Tang¹

^a38 Zheda Road, Zhejiang 310027, China

Abstract

Recent research on automotive driving developed an efficient end-to-end learning mode that directly maps visual input to control commands. However, it models distinct driving variations in a single network, which increases learning complexity and is less adaptive for modular integration. In this paper, we re-investigate human's driving style and propose to learn an intermediate driving intention region to relax difficulties in end-to-end approach. The intention region follows both road structure in image and direction towards goal in public planned routener, which addresses visual variations only and figures out where to go without conventional precise localization. Then the learned visual intention is projected on vehicle local coordinate and fused with reliable obstacle perception to render a navigation score map widely used for motion planning. The core of the proposed system is a weakly-supervised cGAN-LSTM model trained to learn driving intention from human demonstration. The adversarial loss learns from limited demonstration data with one local planned route and enables reasoning of multi-modal behavior with diverse routes while testing. Comprehensive experiments are conducted with real-world datasets. Results show the proposed paradigm can produce more consistent motion commands with human demonstration, and indicates bet-

*Corresponding author

Email addresses: wangyue@iipc.zju.edu.cn (Yue Wang), rxiong@zju.edu.cn (Rong Xiong)

¹State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang, China

²Centre for automotive Systems, The University of Technology, Sydney, Australia

ter reliability and robustness to environment change. Our code is available at <https://github.com/HuifangZJU/visual-navigation>.

1. Introduction

In recent automotive driving research, deep learning tries a revolutionizing way for vehicle control, which directly maps raw pixels from camera image to steering commands in an end-to-end manner[1][2][3]. End-to-end(end2end) approach seeks to avoid steps of building an explicit environment model in conventional approach, which include mapping, localization, planned routening and motion planning, etc. Instead, it optimizes driving variations for perception, planning and reasoning in a single network to maximize overall control performance[1]. In contrast to conventional approach, data for training end2end networks can be collected with relative ease way, i.e., driving around and recording human demonstration control.

However, end2end approach faces with the problem of learning very complex mapping in a single network, which needs intensive supervision to handle huge driving variations. Moreover, it prevents intermediate fusion of visual information with other range finder sensors that help much to avoid obstacles. This raises security concerns, as a small error in vision can induce severe consequence for driving in high frequency control loop. Despite these drawbacks, current research often focuses on end2end setting because it allows to look into plenty of challenges. Some recent methods incorporate additional planned routener as learning input[4][5][6], such as a routed map or a directional instruction, as shown in the top half part of Fig. 1. The planned route captures longer-term motion rules and helps to choose a correct direction upon reaching a fork. It is beneficial and brings performance promotion. Yet the network still lacks transparency of how the planner acts on various driving variations.

In order to address the challenges in end2end approach, we focus on learning an explainable representation following the manner of how human drives with planned routener. Humans may rely on the planned route in public softwares

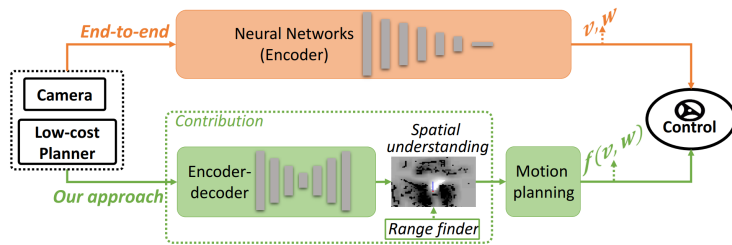


Figure 1: Comparison of proposed paradigm with end-to-end model for automotive driving.

to figure a direction towards goal, then use visual cues like road semantics to reason where to drive. With the goal-directed area in mind, they perform flexible vehicle control in relating to different driving scenarios. The specific control rules may change, e.g., to follow a lane in urban road nets or to mind unexpected obstacles in campuses. However, a goal-directed visual region is always formed based on the local road situation, which keeps an overall sense of driving direction for vehicle control. We denote this region as driving intention.

We think this longer-term driving intention region is effective and informative to improve end2end approach, which addresses only visual variations and solves where to go without conventional precise localization and mapping. Therefore, we define the pipeline as shown in the bottom half part of Fig. 1. An encoder-decoder structure is adopted to learn the aforementioned driving intention from image perception and a planned routener. We follow the work in [5] and resort to the planned route in public available navigation softwares from GPS localization. The driving intention region is then projected onto a local navigation score map with range finder data fused to increase reliability. Such a navigation score map encoded with goal-directed information can be directly used for next motion generation, which is also able to explicitly consider more specified motion variations.

To avoid manual definition and annotation of driving intention on image, we devote to learning from human demonstration. Specifically, human control a vehicle to move and follow a planned route towards goal. Then for each image perception, its traversed area in the near ground satisfies current driving inten-

tion and can be projected on the image plane as supervision. The challenge lies in that the demonstration driving only covers a single direction for each fork and cross, while routes planned to different directions can be provided during test. The driving intention is valid only if it holds rationality in regards to the pair of visual observation and local planned route. Thus, we consider the learning task is not pixel-level imitation but structural reasoning, and develop a weakly-supervised model of cGAN-LSTM network utilizing adversarial loss function. The network learns to generate a ‘fake’ driving intention region which is hard to be distinguished from the ‘real’ region maneuvered by human. Then the generated result is punished as a whole to implicitly learn road semantics, planning intention as well as their correlation. Besides, time continuity is considered with a LSTM unit for performance enhancement. The outline of our method is provided in Fig. 2.

In the experiment, a straightforward motion generation method is implemented in a DWA(Dynamic Window Approach)[7] manner for comparison with end2end system. The proposed pipeline is validated through real-world datasets including previously unseen scenarios to demonstrate generalization performance. Experiments show our approach achieves better performance than state-of-the-art end2end approach and demonstrates improvements in reliability and robustness. To summarize, our main contributions are twofold:

- An innovate learning-based automotive driving system is developed. The system learns from low-cost GPS-level planned routener and images to achieve goal-directed driving intention without precise localization. It eases problem complexity for end2end models and can be efficiently integrated for modular motion planning.
- A weakly supervised and adversarial learning method is developed through learning from demonstration, the core of which is a cGAN-LSTM network trained with limited single-modal demonstration data. The model is enhanced with time continuity and can be generalized to achieve multi-modal behavior when facing new scenarios.

The remainder of the paper is organized as follow: Section 2 reviews the related works on learning-based approaches for automotive driving. Section 3 illustrates details of the proposed system architecture. Section 4 presents the experimental results, and Section 5 draws a conclusion.

85 **2. Related Work**

Conventional pipeline of vehicle control includes mapping, localization, path planning and motion planning. The result of localization, i.e. a reference path or a goal in vehicle local coordinate, and the result of path planning, i.e. a grid map with(out) semantics, are fed into motion planing to generate final control
90 command. The system is thus sensitive to environment change and calls for a lot of work to improve performance on separate modules. In this section, we give a brief review of two learning-based systems aimed at improving the conventional automotive driving, each following a different approach to the system design: the end2end approach and the intermediate recognition approach.

95 *End-to-end approaches.* Recently, end2end method learned from human demonstration becomes popular in automotive driving. The intrinsic merit is that the performance of intermediate stages in conventional system architecture may not be aligned with the ultimate goal, namely, the control of the vehicle. With this idea, [1] firstly proved powerful ability of CNN to steer a vehicle directly from
100 vision input. Codevilla, et,al.[2] then proposed to learn the driving model to compute motion command via conditional imitation learning, which incorporates high-level command input to consider the repeatability of imitation learning. The work in [4] collected control commands from existing local planner (Dynamic Window Approach[7]) and proposed a two-stage approach to relax
105 prior knowledge for localization. This relies on the path-planning results, as the form of navigation is to learn expected motion commands using a residual neural network. The work in [5] adopted 360-degree surround-view cameras along with planned routes information from commercial maps to learn an end2end driving model. Their work has utilized GPS signals as well as public map to

110 generate steering angles and speeds based on a RNN. As reported in their evaluation, it has unavoidably incorporates human intervention. The work in [8] proposed to estimate a variational network to get a full probability distribution over the possible control commands; however, when combined with specific navigation indicators, they still solve an accurate form of certain control command.
115 Compared with this category, our work relies on similar input while achieves intermediate representation in robot local coordinate without relying on precise geometric transformations. We consider the problem space of end2end control learning is more complicated than ours, as the motion states of the vehicle are coupled with the visual understanding.

120 *Intermediate Recognition approaches.* Another group of works focus on CNN based recognition tasks to facilitate automotive driving. One category is following the scene parsing problem to recognize common traffic elements in an image, such as vehicles, pedestrians and cyclists detection [9, 10, 11, 12, 13, 14]. Another idea goes a step further towards vehicle direct usage, including regression of distance to other cars[15], recognition of traversable region[24][16], or road attributes prediction[6], etc. The work in [15] intended to find an intermediate traffic semantics by regressing distances to the lanes as well as other road users. Tang et al, [16] applies mapping techniques to achieve weakly supervised learning of traversable regions. Wang et al, [17] proposed an on-line learning
125 mechanism to deal with the appearance change of traversable region without referring to the massive data. In [18], a model to infer road layout and vehicle relative pose is developed given imagery from on-board cameras, utilizing the public Google Street View and OpenStreetMap. The work in [6] proposed intermediate affordances to facilitate driving, including both vehicle relative pose
130 to the road and recognition of traffic signs, while the human-defined affordances may not be adaptive to different driving scenarios. These works have eventually yielded an intermediate representation for vehicle control, while may still needs further reasoning for motion control. And some of the learned perception result still rely on accurate pose estimation from conventional approach.

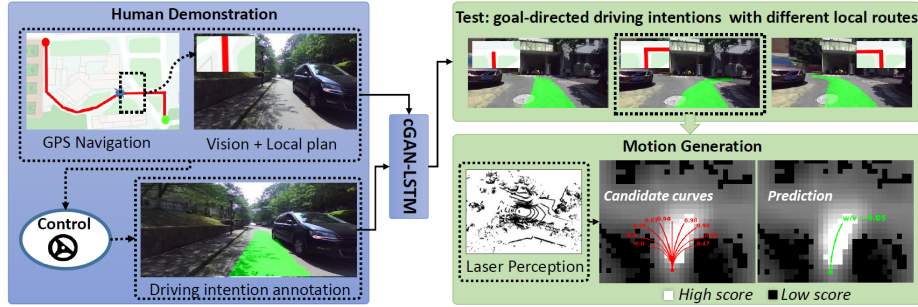


Figure 2: Outline of our approach. A cGAN-LSTM model is utilized to learn driving intention from human demonstration with local planned route. While testing, the model generates corresponding driving intentions following both local routes and road structures. The driving intention is then integrated with concurrently collected laser data and rendered into a navigation score map. Based on this, we generate motion by scoring candidate driving curves.

This section introduces the proposed driving paradigm in detail. The system architecture is shown in Fig. 2, blue box shows the procedure of learning from human demonstration and green boxes show the test application. The core of the system is a cGAN-LSTM network, which takes front-view image and local planned route from public navigation software as input. The model learns to generate a goal-directed driving intention region on the road area to indicate supposed future control. When testing in strange scenarios with different local planned route, the model is able to generate corresponding driving intentions towards different directions. For motion generation, the driving intention is then projected on vehicle near ground and integrated with concurrent laser perception to render a navigation score map. Finally, we implement a straightforward method to generate control command by scoring candidate driving curves. The specific illustrations of each step is provided in the following parts.

3.1. Weakly-Supervised driving intention learning

3.1.1. Network design

We frame the intention learning as a structure reasoning process to follow both road situation in image and planning intention in local planned route.

Since human demonstration only covers a single direction for each fork and cross while routes planned to different goals can be provided during test, we do not treat the learning as a pixel-level imitation and regression. Instead, we adopt the recent GAN[19] model which utilizes adversarial loss to generate and evaluate the network output as a whole. GANs consist of a generator and a discriminator. The fake output from generator is trained to be similar with the real data so as to cheat the discriminator. Thus, the generator eventually learns to produce an overall reasonable result following the distribution in provided dataset. We implement a network structure of cGAN-LSTM specifically for driving intention generation, the model structure is provided in Fig. 3.

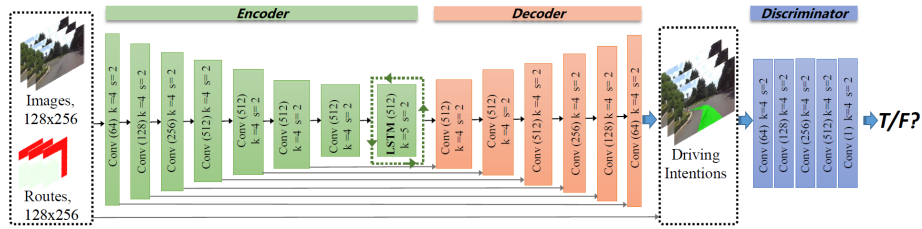


Figure 3: Model architecture of cGAN-LSTM. Front-view images combined with local planned routes are fed into a UNet structure to generate goal-directed driving intentions. The middle of the UNet is inserted with a LSTM-unit to incorporate time continuity. The predictions are then concatenated with input images to go through the discriminator.

The network structure has referred to the work in [21], which follows the design of conditional GAN[20] and utilizes a UNet[22] as generator. As implied, UNet is an encoder-decoder structure with skip connections to keep lower-stage features. We use less layers in our case since the generation task does not need to recover complete textures of the image. Our model treats both the image and the planned route as prior conditions. The two inputs together with the generated driving intention are fed into the discriminator for evaluation. Since the intentions are continuous in both time and space domain during driving, a LSTM(Long Short-term Memory) unit is inserted after the last encoder layer to capture series relation, which at the same time guarantees minimum parameter increase.

Let’s consider $k-1$ steps of the former visual perception, the sequential input images are denoted as $I_{[t-k+1,t]}$ and the corresponding local planned routes are denoted as $R_{[t-k+1,t]}$, for each time t . We expect to learn a visual driving intention towards goal, denoted as V_t , at current image. Thus, the problem is formulated as $G : \{I_{[t-k+1,t]}, R_{[t-k+1,t]}\} \rightarrow V_t$. Since previous approaches have found it beneficial to mix the GAN objective with a more traditional loss, such as L1 distance[21], the objective function is the sum of two weighted loss for the considered k steps:

$$L = \arg \min_G \max_D \sum_t^{t-k+1} L_{cGAN}(G, D) + \lambda L_{L1}(G) \quad (1)$$

where λ is the weight parameter.

The first item is the standard cGAN objective function:

$$L_{cGAN}(G, D) = E_{I_t, R_t, V_t} [\log D(I_t, R_t, V_t)] + E_{I_t, R_t} [\log(1 - D(I_t, R_t, G(I_t, R_t)))] \quad (2)$$

and the second item is a patch-wise L1 distance from generated intention to the provided real driving intention:

$$L_{L1}(G) = E_{I_t, R_t, V_t} [\|V_t - G(I_t, R_t)\|_1] \quad (3)$$

180 The cGAN-LSTM model ensures generated intention to consider both current road structure in image perception and different intentions in local planned routes. It has implicitly learned their inherent correlation by adversarial training and can be generalized to allow for different driving intentions when confronting new scenarios.

185 3.1.2. Data preparation

To achieve the learning of driving intention, we need to provide data of image perception, local planned route as well as annotation of driving intention region. More importantly, their correlation needs to be specifically established. This part illustrates how we make the data ready for training network and
190 performing comprehensive experimental evaluation.

Local planned route We devote to follow human’s manner which uses public navigation softwares to get local planned route with GPS signal. Thus one possibility is to enable the communication between vehicle to such a APP during demonstration. Nevertheless, the interface to real-time synchronized view of local planned route is not make public for research usage. Besides, the model performance under different GPS localization errors needs to be carefully considered and experimentally evaluated. This may lead to substantial workload for on-line data collection and rendering. Therefore, we develop an off-line route rendering method, which makes use of the spatial alignment between public map and vehicle demonstrated trajectories. The procedure is shown in Fig. 4.

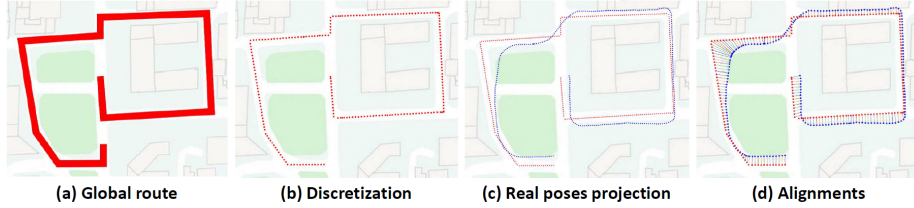


Figure 4: Procedure to get off-line local planned routes.

Given public map data from Baidu Map³, the global route R is annotated in a similar manner to that the navigation software uses, as shown with the red line in Fig. 4(a). Then, the route R is discretized to route points R_d in Fig. 4(b). After human demonstration driving along the global route, we obtain the vehicle poses as shown with blue dots in Fig. 4(c), denoted as T_r .

The spatial alignment from vehicle pose to the routed public map is now the task of aligning two sets of planar points R_d and T_r . Here, we use the DTW (dynamic time warping)[23] algorithm which is commonly adopted in the time domain to warping time series data:

$$DTW(T_r, R_d) = \min \frac{1}{K} \sqrt{\sum_{k=1}^K w_k} \quad (4)$$

³<https://map.baidu.com/>

210 where K is the warpping length, and $w_k = (i, j)_k$ is the warpping weight between T_r to R_d .

The step-by-step optimization objective is:

$$\gamma(i, j) = w(T_r(i), R_d(j)) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (5)$$

where γ is the accumulated series distance. Specifically, a geometric warping criterion is adopted in our scenario, and w is defined as the euclidean distance from projected vehicle pose to the center route point:

$$w(T_r(i), R_d(j)) = \|(T_r(i), R_d(j))\|_2 \quad (6)$$

The aligned result is shown in Fig. 4(d). By assigning the heading direction of each road, we can crop desired local planned route under various experiment settings. The vehicle pose used for data preparation is calculated with the
 215 conventional localization approach in our previous work[16].

Driving intention annotation. We annotate driving intention region from human demonstration, which projects vehicle traversed area under specified local planned route on current image, as shown in Fig. 5.

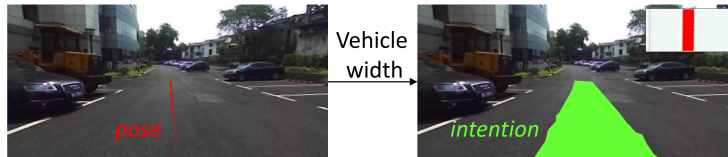


Figure 5: Driving intention annotation. Left: vehicle projected poses; Right: annotated driving intention region.

Human control a vehicle to move and follow a specified route towards goal.
 220 For each image, vehicle future poses in the near ground are first projected on the image plane. Then the poses are dilated with vehicle width to indicated current driving intention. The annotation certainly satisfies planning intention in the planned route and road semantics in image. This idea is similar to the drivable region annotation work in [24]. However, they do not differentiate
 225 region directions and further do not learn its relation with the planned route.

The other region on the image are then similarly labeled with *obstacle* and *unknown*[24] utilizing the projection of concurrently collected laser perception.

3.1.3. Training

We consider four time steps to train the cGAN-LSTM network: 0.9s in the
230 past, 0.6s in the past, 0.3s in the past, and the current frame, similar to the
experiment setting in [5]. As the straight road sections are much longer than
the turning sections, the straight perceptions are down-sampled to around one
sixth to keep a same quantity with that of turning perceptions.

For parameter optimization, we firstly train a basic-model of cGAN without
235 LSTM unit, following the common procedure of one gradient descent on D
and then one step on G . The basic model network is trained with stochastic
gradient descent(SGD) at a learning rate of 0.0002, and momentum parameters
of $\beta_1 = 0.5$, $\beta_2 = 0.999$. We train basic model to 200 epochs with a batch size of
12. Then cGAN-LSTM model is fine-tuned based on the pre-trained parameters
240 of basic model, and the encoder part of UNet is fixed to keep a stability of the
network. The cGAN-LSTM model is trained additionally around 20 epochs. At
inference time, the generator net runs in exactly the same manner as during the
training phase.

3.2. Motion generation with driving intention

245 The learned driving intention preserves a learned region which is highly
adaptive to fuse with other sensors and motion variations to ensure vehicle
safety. Considering vehicle usually runs on smooth roads, we make an assump-
tion that the road area in the near front of vehicle can be modeled with a flat
plane. Thus, driving intention learned to follow road structure in image plane
250 can be projected on robot local coordinate with camera calibration parameters.
Then local driving projection is then integrated with concurrent laser perception
to render a navigation cost map, based on which, a motion generation method
is implemented by scoring candidate driving curves. The procedure is shown in
Fig. 6.

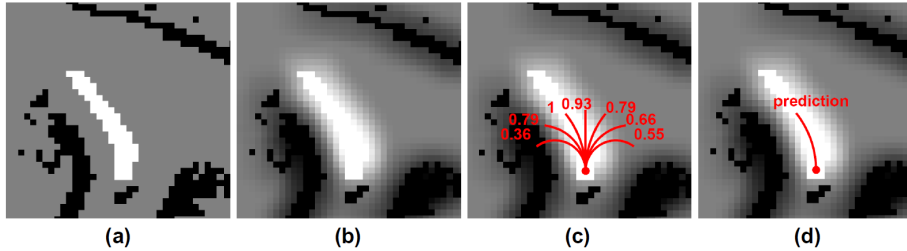


Figure 6: Motion generation with learned driving intention. (a) Projection of visual driving intention with laser perception integrated; (b) Navigation score map modeled with Gaussian kernel; (c) Candidate driving curves with their score labeled; (d) Final control command generated with the driving intention.

255 The white grids($0.5\text{m}\times 0.5\text{m}$) in Fig. 6(a) shows projected driving intention in vehicle local ground. The black grids indicate the obstacle perception for concurrent laser data. To consider neighboring influence, driving intention grids are assigned with positive Gaussian kernels and obstacle grids are assigned with negative Gaussian kernels, which forms the navigation score map used for motion
 260 generation, as shown in Fig. 6(b).

In order to keep the task tractable, we chose to generate motion in an DWA manner, i.e., to produce candidate driving commands and estimate the best one. Following the work in [1][8], we present the steering command as driving curvature, denoted as $\frac{1}{r}$, where r is the turning radius in meters. Since the vehicle
 265 constantly adjust its control based on visual perception, it is reasonable to assume that vehicle keep a uniform motion during a short time clip. Resultantly, vehicle future trajectory can be modeled with a curve in the local navigation score map.

Specifically, the 90° space ahead of vehicle is divided into different number of
 270 direction sections in relating to different requirement for control precision. We then generate a same number of candidate driving curves whose curvatures range evenly form $[-0.2, 0.2]$, as shown in Fig 6(c) which is an example of generating seven candidate curves. Each curve can be estimated with a score based on the navigation score map. Then the final command is the curve with the highest

275 score, indicating direction towards goal, as implied in Fig. 6(d). The motion generation is straightforward while efficient to show its usage and compare with end2end approach, the incorporation of more specified motion variations is the future work.

4. Experiments

280 This section reports the experiment results of the proposed approach, including experiment set up, performance of driving intention generation, and performance of motion generation compared with the state-of-the-art end2end approach.

4.1. Data sets

285 Experiment data is collected with a real vehicle running in our campus, which has been extensively adopted in conventional automotive driving research[25, 26, 27]. The data collection route is shown in the left of Fig.7. Blue line shows the training route with a length of 1.2km. Red line shows the test route with a total length of 4.8km. The overlap sections of the two routes basically runs in a
290 bi-directional manner. The vehicle used for data collection is shown in the right side of Fig.7, which is a four-wheeled mobile vehicle equipped with a ZED stereo Camera, a Velodyne VLP-16 laser scanner and a D-GPS. Only images from the left camera of ZED are used with a resolution of 314×648 pixels. The training data involves 21 times of demonstration driving at different time over three days,
295 covering varying weather/illumination conditions. Each demonstration driving contains ~ 7000 frames of observation. The test data contains ~ 25000 frames of observation.

4.2. Intention generation result

We first present the result of driving intention generation. Since the representation of driving intention region is a novel contribution in our work,
300 we propose two new criteria along with the common used IOU for quantitative evaluation.

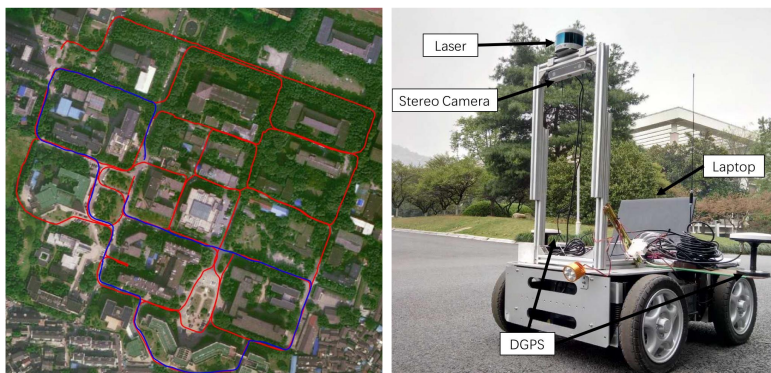


Figure 7: Data collection route and experiment vehicle.

Evaluation Metrics. As shown in Fig. 8, dark gray represents the demonstrated human driving and light gray grids shows the predicted intention. Green dots and red dots show the center line of the two intentions respectively. Based on these representation, we employ three criteria for visual goal-directed intention evaluation. The first is IOU, intersection over union between prediction and human demonstration. The second is *cover_rate*, ratio of predicted central line falling inside demonstrated intention region, which is measured as the percentage of green dots falling in the dark gray region. And the last is Δyaw , angle difference of driving directions between prediction and human demonstration in image plane. The driving direction is measured as the angle between central line and horizontal line.

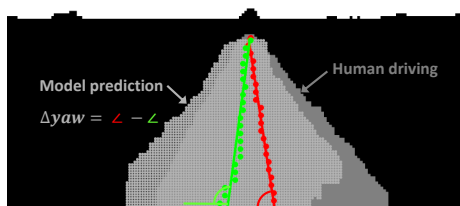


Figure 8: Illustration of visual evaluation. The dark gray shows the future control of human demonstration and the light gray grids imply model prediction. Red dots and green dots show the center lines of human driving and model prediction respectively, which have been both downsampled for better visualization.

4.2.1. Network performance

We test the visual result with two models of *cgan_basic* and *cgan_lstm*.
315 The *cgan_basic* model does not include a LSTM unit. The result is presented
in Table 1.

Table 1: PERFORMANCE OF DRIVING INTENTION GENERATION

Model	iou%	cover_rate%	Δyaw
<i>cgan_basic</i>	62.06	95.9	13.71
<i>cgan_lstm</i>	78.01	98.9	9.05

From the table we can see, *cgan_lstm* outperforms *cgan_basic* for all the
three criteria. The LSTM unit has shown its effectiveness to model time con-
tinuity and shape the driving intentions. The *cover_rate* of both models has
320 exceeded 95%, which implies the predicted intention basically follows the road
structure, as the human demonstration region is certainly on the road area.
The model of *cgan_lstm* has achieved more improvement on the IOU which
is 78.01%, and the Δyaw which is around 9 degrees. A higher IOU value im-
plies more similarity to human demonstration. However, we do not seek a
325 one-hundred-percent result, since the driving intention is not defined with an
explicit boundary and there can be some pixel-level shape variations. In this
situation, Δyaw shows more importance for direction measurement. We con-
sider 9-degree is a favorable value, since the case in Fig. 8 has a 11-degree
different in Δyaw , where the predicted intention is structural close to human
330 demonstration. More qualitative results from *cgan_lstm* are presented in Fig.
9.

The first row of Fig. 9 presents some cases for driving on straight roads. The
straight road is relatively easy task for intention generation. And the mainly
challenge lies on the recognition of obstacles, as shown in the last figure of
335 this row. It provides an example when confronting with a pedestrian which is

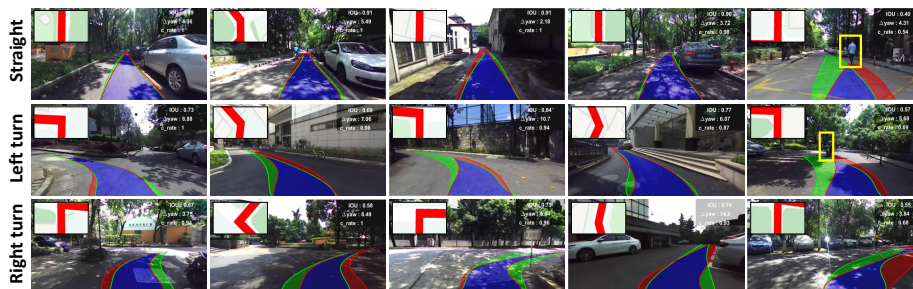


Figure 9: Visual results from cGAN-LSTM. Green, red and blue colors represent prediction, human demonstration and their intersection respectively. The local planned routes are shown in the top left corner of the image and the evaluation figures are annotated in the top right corner of the image. The intention boundaries are fitted with splines in these figures.

annotated with a yellow box. The generated intention has turned to a more central direction than that of human demonstration. Thus, it causes lower criteria values.

The last two rows show some cases for left turns and right turns respectively. The proposed model has generated correct driving intentions following various local road situations. For the turning classes, we see more shape changes to that of human demonstration. The result satisfies the design of adversarial loss, which focus more on the structural rationality rather than pixel-level imitation. Besides, the yellow box in the second row indicates a cyclist where the predicted intention has also avoided. Therefore, the proposed model seems to show an ability of avoiding dynamic obstacles on the road without explicit annotation.

4.2.2. Robustness to Localization errors

We assume to utilize public navigation softwares from GPS signal. As the GPS signal commonly provides rough localization results, this section discusses the model performance under different localization errors. In order to achieve it, we have randomly added horizontal and vertical offsets when rendering local planned route, as illustrated in Section 3.1.2. The random offset goes into three levels, easy, moderate and hard, with each level corresponds to a localization error of $0m \sim 1m$, $1m \sim 2.5m$, and $2.5m \sim 5m$. The result is presented in

Table 2

Table 2: ROBUSTNESS TO LOCALIZATION ERRORS

Model	minor(0m~1m)			moderate(1m~2.5m)			hard(2.5m~5m)		
	iou	c%	Δyaw	iou	c%	Δyaw	iou	c%	Δyaw
<i>basic</i>	61.8	96.2	13.9	61.6	95.8	14.2	61.5	95.9	14.2
<i>lstm</i>	76.5	98.4	9.9	76.1	98.3	10.3	75.9	98.4	10.5

abbreviations: *basic*(*cgan_basic*), *lstm*(*cgan_lstm*), *c%*(*c_rate%*)

355

As can be seen in the table, the two models have basically achieved a stable performance given different level of route offsets. To compare with the result from center-view routes in Table 1, only $\Delta yaw/deg$ has a slight increase with the growing of localization errors, while the other two criteria have remained in a similar value to the previous result. The model has shown robustness to potential localization errors. Moreover, it verifies the proposed pipeline can get a local reference path towards goal without common requirement to precise geometric localization. Some visual result from *cgan_lstm* with different level of route offsets are shown in Fig. 10. The offset in local planned route does not have a severe influence for intention generation, as the annotated route clearly conveys an overall sense of direction within visible range.

360

365

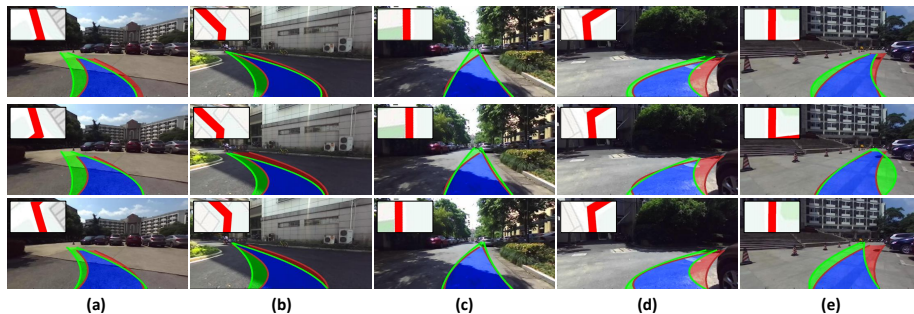


Figure 10: Robustness to localization errors. From top to bottom: minor(0m~1m), moderate(1m~2.5m) and hard(2.5m~5m) errors. Green, red and blue colors represent the prediction, human demonstration and their intersection respectively.

4.2.3. Discussion: multi-modal behavior

Annotation from human demonstration only validates a single driving intention with pre-defined local planned route, while multi-modal behaviors are presented when approaching intersections and open areas. For a further qualitative evaluation, three fake planned routes are made to test the model, which are intuitively viewed as {go-straight, turn-right, turn-left}. These planned routes are all used to generate driving intentions on the test image perceptions, and Fig. 11 provides some visual result from a direct network output.

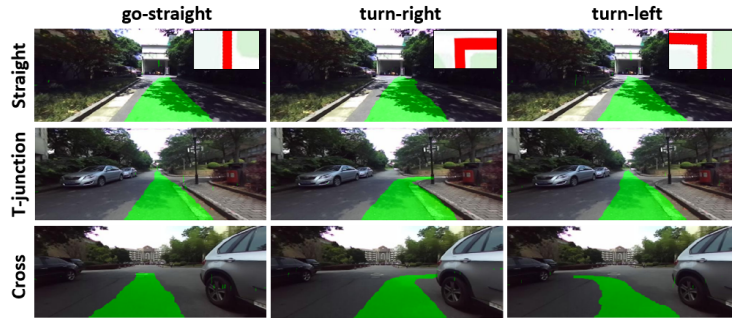


Figure 11: Multi-modal driving behaviors with different road types. From top to bottom: one-way straight road, T-junction road, and cross area. Green color shows the original output from the network.

The three rows show the road types of straight, T-junction and cross respectively. The local planned routes are shown in the top right corner on the first row. For the first two classes, there are local planned routes that may not be allowed on current road situation. In this case, the model can still generate reasonable intentions following road structure. While for the last class, where all planned routes can be performed, the model generates driving intentions accordingly. Here, we did not use a spline to fit the intention boundary, which better shows the direction difference on network outputs for different planned routes.

Fig. 12 has specifically presented a group of images when facing a moving car

with different planned routes. The result is visually compared with the model
 385 of *pix2pix*[21] which does not include planned routes for intention generation.
 When there are dynamic obstacles appear in front of the road, the proposed
 model accordingly adjust the intention generation, which further implies the
 weakly-supervised model has learned to avoid obstacles without explicitly an-
 notation.

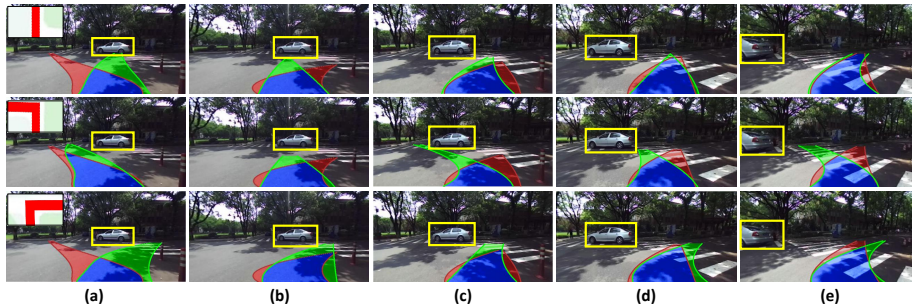


Figure 12: Multi-modal driving behaviors when facing a moving car. Green, red and blue show the result from cGAN-LSTM, pix2pix and their intention respectively. The intention boundaries are fitted with splines in these figures.

390 4.3. Motion generation result

To show the effectiveness of proposed driving intention on motion generation,
 a straightforward method is performed by scoring candidate driving curves. For
 comparison to end2end approach, we implement the network structure in [5],
 which generates direct control commands with public planned route. Model
 395 in [5] outputs driving commands as velocity and angular speed. We calculate
 driving curvature as $c_t = \frac{w_t}{c_t}$, for each time t referring to [8]. As for the ‘ground
 truth’ from human demonstration, the actual trajectory curvature is calculated
 for comparison.

Evaluation Metrics. To quantitatively evaluate the system performance,
 400 we compute the motion prediction accuracy for different control precisions. The
 definition of true positive is illustrated in Fig. 13.

The black dials implies different motion resolutions, for which the proposed
 method generates a same number of candidate motion commands, as shown

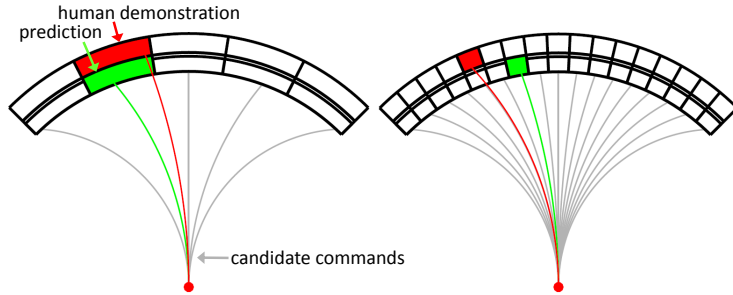


Figure 13: Evaluation criterion for motion generation. Gray curves show the candidate driving commands in relation to the control precision. Green curve indicates the command with a highest score in the navigation score map. And red color shows the ground truth maneuvered by human demonstration.

with the gray curves. Green grid indicates final model prediction and red grid
 405 indicates human demonstration. Then the prediction accuracy is measured with
 the grid distance Δg from model prediction to human demonstration:

$$accuracy = \frac{\sum_{t=0}^T [|\text{pred}_t - \text{human}_t| \leq \Delta g]}{T}$$

where T is the total step of test dataset. $[\cdot]$ equals to 1 if the formulation
 inside is true, otherwise equals to 0. For comprehensive evaluation, we provide
 quantitative results of three settings of $\Delta g = \{0, 1, 2\}$ under control resolutions
 410 for 3 to 23. The resolution of 3 means the prediction only distinguish the
 direction from right, left to straight. And a resolution of 23 represents a control
 precision of less than 4° .

4.3.1. Model performance

We have tested our two models of *cgan_basic* and *cgan_lstm* to compare
 415 with end2end[5] method. The results with different motion resolutions
 are shown in Fig. 14.

The three figures present accuracy on complete test route, long straight test
 route, and turning route respectively. For an overall accuracy measurement,
cgan_lstm achieves better performance than end2end method under most set-
 420 tings and shows better robustness to different control resolutions. When the

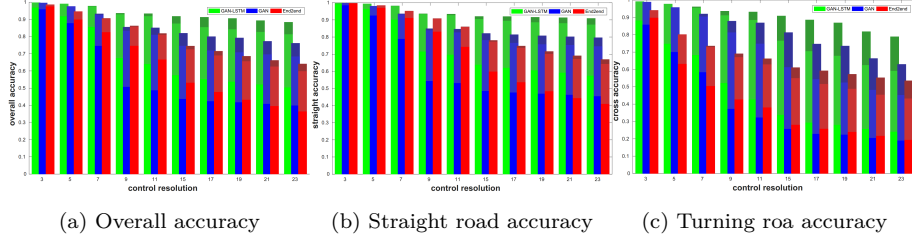


Figure 14: Motion prediction performance. Green, blue, and red represent the performance of *cgan_lstm*, *cgan_basic* and *end2end* respectively. For each individual model result, color from light to dark represent a grid distance 0,1,2 for Δg .

control resolution goes up, *cgan_lstm* also generates more candidate curves for a careful search to better fit for the driving intention. This shows the benefit to separate driving variations on perception from that on motion in the proposed system. The visual driving intention is highly adaptive to different motion requirements, regardless of specific control precision in the demonstration driving, as the intention region explicitly considers local road situation and leaves a room for a second motion module to follow.

In contrast, *end2end* approach learns a numerical mapping from visual input to output. Thus, *end2end* outputs a same result for different control resolutions, and shows steady decrease when the evaluation gets more strict. For the performance on separate straight class and turning class, advantage of the proposed model is more significant for the turning road sections which is more challenging due to the complex road situation. A more intuitive comparison of error rates along test route between *cgan_lstm* and *end2end* are provided in Fig. 15.

As the figure shows, *end2end* method have more errors when vehicle approaching turning road sections. We consider the reason may be the numerical difference of demonstration control when facing a same turning class. A big turn requires a small control curvature and a small turn may require a big control curvature. Besides, human may change the driving command during turning for obstacle avoidance. This also increases the intra-class variation for similar perceptions. Thus, it may be difficult for *end2end* training to consider

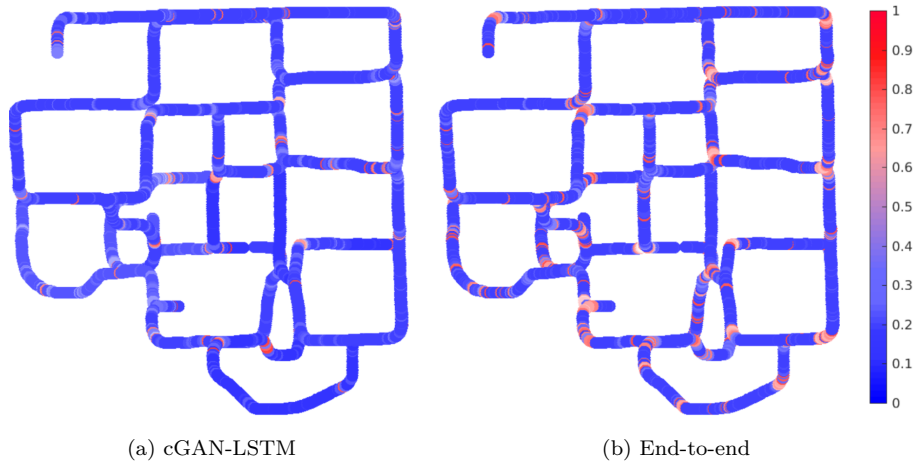


Figure 15: Error rate density along the test route. The results come from a control resolution of 7-grid. Different colors represent different error levels, and the blue the better.

both variations in perception and planning. In contrast, for the long straight road section, most commands stay near-linear and unchanged for a long time, which can be efficiently learned by both models. Some visual result from the proposed system is shown in Fig. 16.

4.3.2. Discussion: robustness to time delay

For the design of end2end approach, motion generation is correlated with environment understanding. This makes the system has a strong reliance to real-time vision prediction. However, the vision processing alone is prone to be disturbed and the GPS signal for local planned route can be lost due to occlusions. Thus, it can be a critical ability to generate valid motion when visual result is delayed. In this section, we investigate the motion generation ability when there are different level of time delays for visual prediction. The results are shown in Fig.17 and Fig. 18 for end2end method and the proposed method respectively.

The figures show sequential prediction curvatures along a section of test route. The discrete dots represent actual data points, which are fitted with smooth lines to indicate variation trend. End2end approach outputs direct

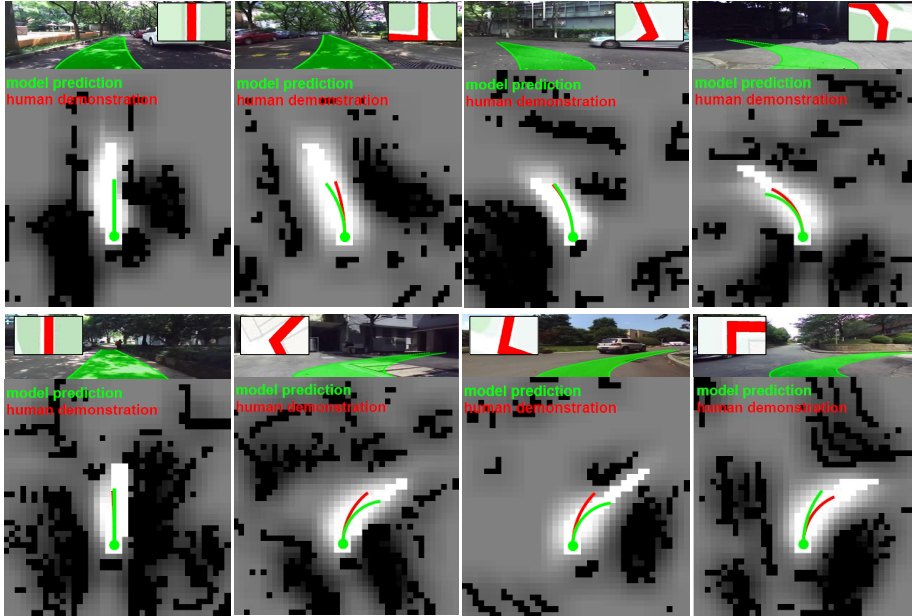


Figure 16: Motion generation result compared with human demonstration. For each case, top figure shows the visual result and bottom figure shows the generated motion on navigation score map. Green and red colors represent the model prediction and human demonstration respectively.

driving commands without intermediate knowledge retainment. Thus, during
 460 time delay, it has no choice but to keep the former motion command without
 human intervention. This is reflected on the extended dashed lines for motion
 generation in Fig. 18.

Nevertheless, for the proposed method, driving intention prediction is sep-
 arated with the motion generation. The learned intention has a certain area
 465 on the ground plane, which can be used for motion reference in multiple steps.
 Thus, driving intention in one frame can be efficiently integrated with follow-
 ing obstacle perception, and be used to render a new navigation score map for
 motion generation. As can be inferred in Fig. 17, *cgan_lstm* model is able to
 generate new commands when there are temporary missing of vision predictions.
 470 This is reflected on the similar level of dispersion on the data points. Compared
 with the real-time visual prediction, most prediction errors for *cgan_lstm* ap-

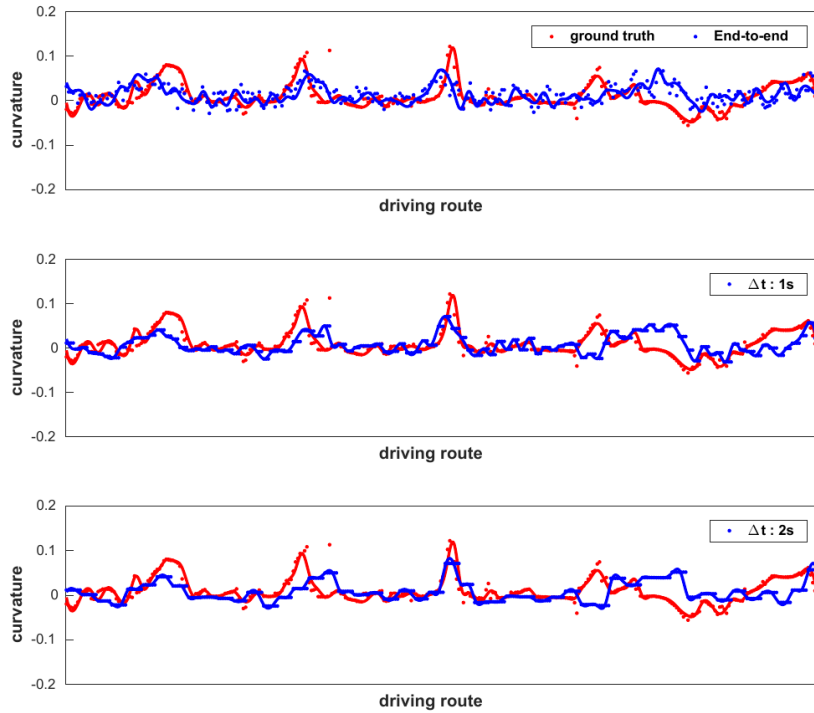


Figure 17: Comparison of end2end method with human demonstration when visual prediction is delayed. From top to bottom: no time delay, 1s delay, and 2s delay. Blue color indicates the model prediction and red color denotes human demonstration. The dots denoted actual data points, which are fitted to a smooth line to show the prediction tendency.

pear as big curvatures. It is caused by planning circle actions for vehicle when it almost passed the retained intention area. In this case, the area around vehicle position has the biggest scores and the model tend to give the largest curvature to stay nearby. Some visual results for a delay of 1.8s are shown in Fig.19

For the proposed approach, the first frame generates a visual driving intention according to local planned route, which is rendered into a navigation score map with laser perception integrated. Then for the following frames, the navigation score maps are integrated from the driving intention in the first frame and the laser perceptions in the subsequent frames. To indicate vehicle movements inside the retained driving intention area, we also plot vehicle previous poses with discrete red dots as shown in the figure. The short transformation of

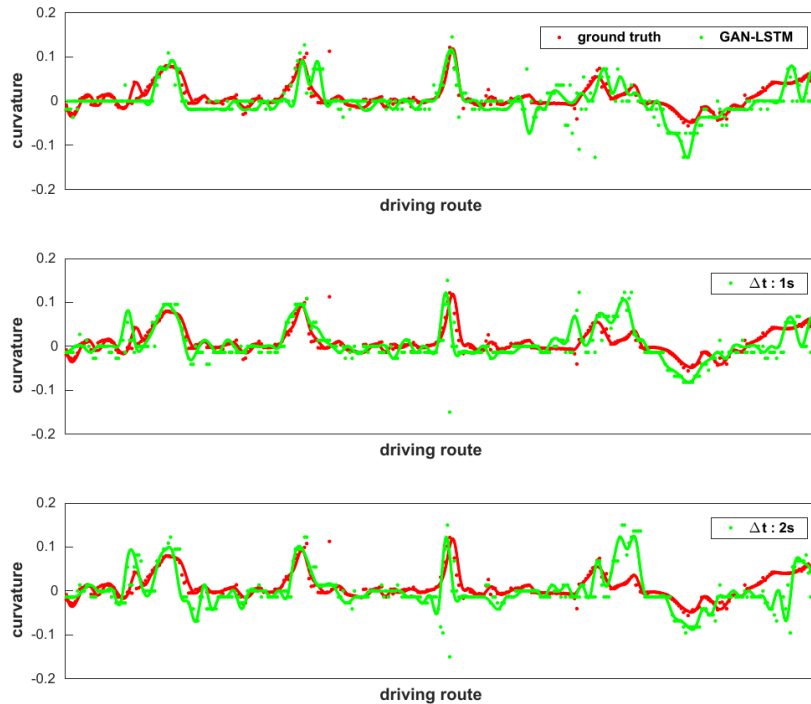


Figure 18: Comparison of proposed method with human demonstration when visual prediction is delayed. From top to bottom: no time delay, 1s delay, and 2s delay. Green color indicates the model prediction and red color denotes human demonstration. The dots denoted actual data points, which are fitted to a smooth line to show the prediction tendency.

local coordinates can be approximately estimated by map registration of laser perception or the integration of motion command.

485 In summary, the proposed method separates driving variations of visual understanding to motion generation. Compared with end2end approach, it demonstrates better prediction performance and robustness to time delay.

5. CONCLUSIONS

490 In this paper, we developed an innovative learning model for automotive driving research with an intermediate driving intention learned from image perception and publicly available planned routener. The driving intention can be efficiently encoded into a navigation score map which is able to directly gener-

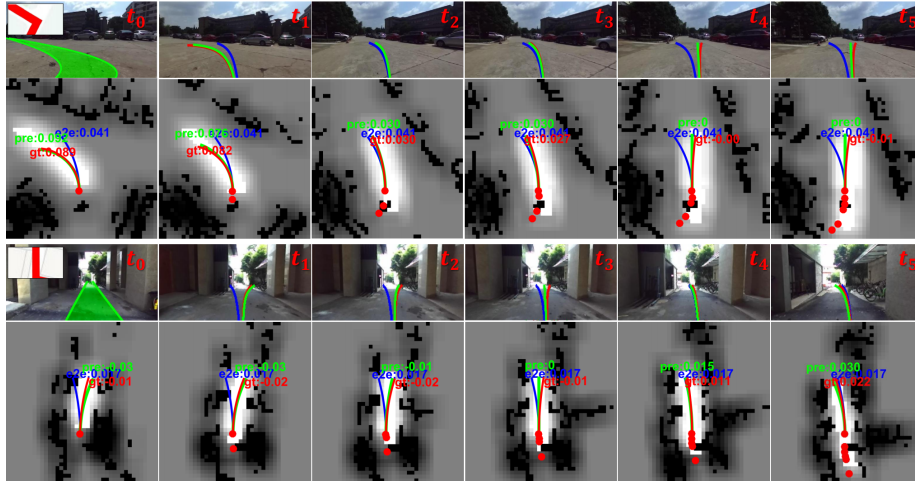


Figure 19: Robustness to time delay. The time difference between consecutive frames is $\delta t = 0.3s$. Green, blue, and red colors indicate motion generation results of cGAN-LSTM, end2end, and human demonstration respectively. The red points in the navigation score map indicates vechile previous pose.

ate motion. In this way, the variations on motion is separated with that on visual understanding, which inceases modular flexibility and reliability compared
 495 with end2end approach. The key of the system is a cGAN network inserted with LSTM unit to learn from human demonstration. The adversarial loss enables a weakly-supervised training manner, which leverages limited single-modal demonstration data to achieve generalization on multi-modal behavior in strange scenarios. Experiments indicates driving intention generation is robust
 500 to errors on vehicle global pose and shows more attributes of reliability and adaptability for consideration of real applications. Our future work will incorporate more detailed obstacle modeling modules, which further narrows the gap for real application.

ACKNOWLEDGMENT

505 This work was supported in part by the National Key R&D Program of China (2017YFB1300400), in part by the National Nature Science Foundation of China (U1609210).

References

- [1] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, K. Zieba, End to
510 end learning for self-driving cars.
- [2] F. Codevilla, M. Müller, A. Dosovitskiy, A. López, V. Koltun, End-to-
end driving via conditional imitation learning.
- [3] H. Xu, G. Yang, F. Yu, T. Darrell, End-to-end learning of driving models
from large-scale video datasets, in: IEEE Conference on Computer Vision
515 and Pattern Recognition, 2017.
- [4] W. Gao, D. Hsu, W. S. Lee, S. Shen, K. Subramanian, Intention-net: In-
tegrating planning and deep learning for goal-directed autonomous naviga-
tion, arXiv preprint arXiv:1710.05627.
- [5] S. Hecker, D. Dai, L. Van Gool, End-to-end learning of driving models
520 with surround-view cameras and route planners, in: Proceedings of the
European Conference on Computer Vision (ECCV), 2018, pp. 435–453.
- [6] A. Sauer, N. Savinov, A. Geiger, Conditional affordance learning for driving
in urban environments, arXiv preprint arXiv:1806.06498.
- [7] D. Fox, W. Burgard, S. Thrun, The dynamic window approach to collision
525 avoidance, IEEE Robotics & Automation Magazine 4 (1) (1997) 23–33.
- [8] A. Amini, G. Rosman, S. Karaman, D. Rus, Variational end-to-end naviga-
tion and localization, arXiv preprint arXiv:1811.10119.
- [9] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv
preprint arXiv:1804.02767.
- 530 [10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg,
Ssd: Single shot multibox detector, in: European conference on computer
vision, Springer, 2016, pp. 21–37.

- [11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE transactions on pattern analysis and machine intelligence* 40 (4) (2018) 834–848.
- 535
- [12] Q. Luo, H. Ma, Y. Wang, L. Tang, R. Xiong, 3d-ssd: Learning hierarchical features from rgb-d images for amodal 3d object detection, arXiv preprint arXiv:1711.00238.
- [13] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast r-cnn for pedestrian detection, *IEEE transactions on Multimedia* 20 (4) (2018) 985–996.
- 540
- [14] F. Yang, W. Choi, Y. Lin, Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2129–2137.
- 545
- [15] C. Chen, A. Seff, A. Kornhauser, J. Xiao, Deepdriving: Learning affordance for direct perception in autonomous driving, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2722–2730.
- [16] L. Tang, X. Ding, H. Yin, Y. Wang, R. Xiong, From one to many: Unsupervised traversable area segmentation in off-road environment, in: *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, 2017, pp. 787–792.
- 550
- [17] Y. Wang, Y. Liu, Y. Liao, R. Xiong, Scalable learning framework for traversable region detection fusing with appearance and geometrical information, *IEEE Transactions on Intelligent Transportation Systems*.
- 555
- [18] A. Seff, J. Xiao, Learning from maps: Visual common sense for autonomous driving, arXiv preprint arXiv:1611.08583.

- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,
560 S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
- [20] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784.
- [21] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with
565 conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
- [22] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp.
570 234–241.
- [23] D. J. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series., in: KDD workshop, Vol. 10, Seattle, WA, 1994, pp. 359–370.
- [24] D. Barnes, W. Maddern, I. Posner, Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy, in: Robotics and Automation (ICRA), 2017 IEEE International Conference on, IEEE, 2017,
575 pp. 203–210.
- [25] X. Ding, Y. Wang, D. Li, L. Tang, H. Yin, R. Xiong, Laser map aided visual inertial localization in changing environment, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE,
580 2018, pp. 4794–4801.
- [26] L. Tang, Y. Wang, X. Ding, H. Yin, R. Xiong, S. Huang, Topological local-metric framework for mobile robots navigation: a long term perspective, *Autonomous Robots* 43 (1) (2019) 197–211.
- [27] H. Yin, Y. Wang, L. Tang, X. Ding, R. Xiong, Locnet: Global localization
585 in 3d point clouds for mobile robots, in: Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 2018, pp. 26–30.