Supplementary Material for End-to-End Urban Driving by Imitating a Reinforcement Learning Coach

Zhejun Zhang[1], Alexander Liniger[1], Dengxin Dai[1,2], Fisher Yu[1] and Luc Van Gool[1,3]
[1]Computer Vision Lab, ETH Zürich, [2]MPI for Informatics, [3]PSI, KU Leuven
{zhejun.zhang,alex.liniger,dai,vangool}@vision.ee.ethz.ch, i@yf.io

1. Summary

In this document, we provide (1) an overview of supplementary videos and codes, (2) implementation details of the RL experts and the IL agents, (3) details regarding benchmarks, and (4) additional experimental results.

2. Other Supplementary Materials

2.1. Videos

To investigate how different agents actually drive, we provide three videos. **roach.mp4** shows the driving performance of Roach, and highlights that it has a natural driving style and that it can handle complex traffic scenes. In **autopilot.mp4** we demonstrate the rule-based CARLA Autopilot. This agent uses unnatural brake actuation, i.e. it only uses emergency braking. Further, this video also highlights that in dense traffic, the rule-based agent can get stuck due to conservative danger predictions. For more details about the Autopilot and changes we made see Section 3.3. Finally, in **il_agent.mp4** we demonstrate our best roach-supervised IL agent, showing that the agent can handle complex traffic scenes but also highlighting failure cases. In detail:

- **roach.mp4** is an *uncut* evaluation run recorded from Roach driving in Town03 (LeaderBoard-busy under dynamic weather). This video demonstrates the natural driving style of Roach even in challenging situations such as US-style traffic lights, unprotected left turns, roundabouts and stop signs.
- **autopilot.mp4** is an *uncut* evaluation run recorded from Autopilot driving in Town02 (NoCrash-dense, new town & new weather). This video demonstrates the over-conservative behavior of the Autopilot while driving in dense traffic. This often leads to red light infractions and blockage (both are present in the video).
- **il_agent.mp4** is a *highlight* video recorded from our best roach-supervised IL agent $\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{F}}(\mathrm{c})$. This video includes multiple challenging situations often encountered during urban driving, such as EU and US-style junctions, unprotected left turns, roundabouts and reacting to pedestrians walking into the street. Furthermore, we highlight some of the failure modes of our camera-based IL agent, including not coming to a full stop for stop signs, collisions at overcrowded intersections and oscillation in the steering if the lane markings are not visible due to sun glare. We believe that including memory in the IL agent policy can help in most of these issues, due to a better understanding of the ego-motion (stop sign and oscillations) and other agents' motion (collisions).

2.2. Code

To reproduce our results, we provide four python scripts:

- *train_rl.py* for training Roach.
- *train_il.py* for training DA-RB (CILRS + DAGGER).
- *benchmark.py* for benchmarking agents.
- *data_collect.py* for collecting on/off-policy data.

It is recommended to run our scripts through bash files contained in the folder *run*. All configurations are in the folder *config*. Our repository is composed of two modules:

- *carla_gym*, a versatile OpenAI gym [1] environment for CARLA. It allows not only RL training with synchronized rollouts, but also data collection and evaluation. The environment is configurable in terms of weather, number of background pedestrians and vehicles, benchmarks, terminal conditions, sensors, rewards for the ego-vehicle and etc.
- *agents*, which includes our implementation of Autopilot (in *agents/expert*), Roach (in *agents/rl_birdview*) and DA-RB (in *agents/cilrs*).

2.3. Rendering issues

As illustrated in Fig. 1, on CARLA 0.9.11 reflections from after-rain puddles are sometimes wrongly rendered as black pixels. When the black pixels are accumulated, for example in the middle of Fig. 1a, they are often recognized as obstacles by the camera-based agents. Since this kind of reflection only appears under the testing weather but not under the training weather, generalizing to testing weather is

exceptionally hard on CARLA 0.9.11 for the camera-based end-to-end IL agents.

3. Implementation Details

3.1. Roach

The network architecture of Roach can be found in Table 3 and the hyper-parameter values are listed in Table 5.

BEV: Cyclists and pedestrians are rendered larger than their actual sizes, this allows us to use a smaller image encoder with less parameters for Roach. Additionally, increasing the size naturally adds some caution when dealing with these vulnerable road users.

Update: The policy network and the value network are updated together using one Adam optimizer with an initial learning rate of 1e-5. The learning rate is scheduled based on the empirical KL-divergence between the policy before and after the update. If the KL-divergence is too large after an update epoch, the update phase will be interrupted and a new rollout phase will start. Furthermore, a patience counter will be increased by one and the learning rate will be reduced once the patience counter reaches a threshold.

Rollout: Before each update phase a fixed-size buffer will be filled with trajectories collected on six CARLA servers, each corresponds to one of the six LeaderBoard maps (Town1-6).

Terminal Condition: An episode is terminated if and only if one of the following event happens.

- Run red light: examination code taken from the public repository of LeaderBoard. Terminal reward: $-1-s$.
- Run stop sign: examination code taken from the public repository of LeaderBoard. Terminal reward: $-1-s$.
- Collision registered by CARLA: based on the physics engine. Any collision with intensity larger than 0 is considered. Terminal reward: $-1-s$.
- Collision detected by bounding box overlapping in the BEV. Terminal reward: $-1-s$.
- Route deviation: triggered if the lateral distance to the lane centerline of the desired route is larger than 3.5 meters. Terminal reward: -1.
- Blocked: speed of the ego-vehicle is slower than 0.1 m/s for more than 90 consecutive seconds. Terminal reward: -1.

with s is the ego-vehicle's speed. The terminal reward is the reward given to the very last observation/action pair before the termination. For non-terminal samples, the terminal reward is 0.

Reward Shaping: The reward is the sum of the following components.

- r_speed: equals to $1.0 - |s - s_{\text{desired}}|/s_{\text{max}}$, where s is the measured speed of the ego-vehicle, s_{max} is the maximum speed and s_{desired} is the desired speed. We use a constant maximum speed $s_{\text{max}} = 6$ m/s. The desired speed is a variable and is explained below.
- r_position: equals to $-0.5\Delta_{\text{p}}$, where Δ_{p} is the lateral distance (in meters) between the ego-vehicle's center and the center line of the desired route.
- r_rotation: equals to $-\Delta_{\text{r}}$, where Δ_{r} is the absolute value of the angular difference (in radians) between the ego-vehicle's heading and the heading of the center line of the desired route.
- r_action: equals to -0.1 if the current steering differs more than 0.01 from the steering applied in the previous step.
- r_terminal: the aforementioned terminal reward.

The desired speed, as proposed in [4], depends on rule-based obstacle detections. If there's no obstacle detected, the desired speed equals to the maximum speed. If an obstacle is detected, based on the distance to the obstacle the desired speed is linearly decreased to 0. As obstacle detector we use the hazard detection of Autopilot (cf. Section 3.3). As a dense and informative reward, r_speed helps substantially to train our Roach and the camera-based end-to-end RL agent [4]. However, using rule-based obstacle detections inevitable introduces bias, the trained RL agent can be over-aggressive or over-conservative depending on the false positive and false negative rate of the detector. For example, during multi-lane freeway driving, our Roach decelerates for vehicles on the neighbouring lanes because those vehicles are detected as obstacles during training. Another example, Roach tends to collide after a right turn, this is related to the sector shaped (around 40 degrees) detection area used by the obstacle detection; vehicles and pedestrians on the right are not covered in the detection area. To further improve the performance of Roach, this r_speed should be modified, either using a better obstacle detector, or completely remove the rule-based obstacle detection, and build a less artificial reward based on simulation states.

Mode of Beta Distribution: We take the distribution mode as a deterministic output. The mode of the Beta distribution $\mathcal{B}(\alpha, \beta)$ is defined as

$$\mathrm{M} = \begin{cases} \frac{\alpha-1}{\alpha+\beta-2} & \text{if } \alpha > 1, \beta > 1 \\ 0 & \text{if } \alpha \leq 1, \beta > 1 \\ 1 & \text{if } \alpha > 1, \beta \leq 1 \\ \text{bimodal } \{0, 1\} & \text{if } \alpha < 1, \beta < 1 \\ \text{any value in } [0, 1] & \text{if } \alpha = 1, \beta = 1 \end{cases} \quad (1)$$

For a natural driving behavior, we use the mean $\frac{\alpha}{\alpha+\beta}$ as the deterministic output when the mode is not uniquely defined, i.e. when $\alpha < 1, \beta < 1$ or $\alpha = 1, \beta = 1$.

(a) Reflections from after-rain puddles in fornt of the ego-vehicle are incorrectly rendered as black pixels.

(b) Reflections are correctly rendered if the puddle is not directly in front of the ego-vehicle.

Figure 1: **Rendering issue of CARLA 0.9.11 running on Ubuntu with OpenGL.**

3.2. IL Agent Supervised by Roach

The network architecture of our IL agent is found in Table 4 and the hyper-parameter values are listed in Table 6.

Network Architecture: We use six branches: turning left, turning right and going straight at the junction, following lane, changing to the left lane and changing to the right lane.

Off-policy Data Collection: Following CILRS [2], triangular perturbations on actions are applied while collecting the off-policy expert dataset to alleviate the covariate shift. The off-policy dataset for NoCrash includes 80 episodes and for LeaderBoard it includes 160 episodes. Each episode is at most 300 seconds and at least 30 seconds long. The episode will be terminated if the expert violates any traffic rules, including red light infractions, stop sign infractions and collisions. In such a case, we remove the last 30 seconds of that episode so as to ensure that the off-policy dataset includes only correct demonstrations. Data is not collected using the given training routes but from randomly spawned start and target locations.

On-policy Data Collection: We follow DA-RB [3] for DAGGER with critical state sampling and replay buffer. New DAGGER-data will replace the old data in the replay buffer, while the buffer size is fixed. The same number of frames are contained in the replay buffer as in the off-policy dataset. At each DAGGER iteration, around 15-25% of the replay buffer is filled with new DAGGER-data, whereas at least 20% of the replay buffer is filled with off-policy data. Identical to the off-policy data collection, we use randomly spawned start and target locations while collecting DAGGER datasets. Following DA-RB, we did not use a mixed agent/expert policy to collect DAGGER datasets. However, our code allows this kind of rollout for DAGGER.

Training Details: Since we take the ResNet-34 pre-trained on ImageNet, the input image is normalized as suggested. In case the IL agent uses a distributional action head and/or a value head, the corresponding weights will be loaded from the Roach model at the first training iteration (the behavior cloning iteration). At each DAGGER iteration, the training continuous from the last epoch of the previous DAGGER iteration. We apply image augmentations using code modified from CILRS. The image augmentation methods are applied in random order and include Gaussian blur, additive Gaussian noise, coarse and block-wise dropouts, additive and multiplicative noise to each channel, randomized contrast and grayscale. All models are trained for 25 epochs using the ADAM optimizer with an initial learning rate of 2e-4. The learning rate is halved if the validation loss has not decreased for more than 5 epochs.

3.3. Autopilot

The CARLA Autopilot (also called roaming agent) is a simple but effective automated expert based on hand-crafted rules and ground-truth simulation states. The Autopilot is composed of two PID controllers for trajectory tracking and hazard detectors for emergency brake. Hazards include

- pedestrians/vehicles detected ahead,
- red lights/stop sings detected ahead,
- negative ego-vehicle speed, for handling slopes.

Locations and states of pedestrians, vehicles, red lights and stop signs are provided as ground-truth by the CARLA API. If any hazard appears in a trigger area ahead of the ego-vehicle, Autopilot will make an emergency brake with $throttle = 0$, $steering = 0$, $brake = 1$. If no hazard is detected, the ego-vehicle will follow the desired path using two PID controllers, one for speed and one for steering control. The PID controller takes as input the location, rotation and speed of the ego-vehicle and the desired route specified as dense (1 meter interval) waypoints. The speed PID yields $throttle \in [0, 1]$ and the steering PID yields $steering \in [-1, 1]$. We tuned the parameters for PID controllers and hazard detectors manually, such that the Autopilot is a strong baseline. The target speed is 6 m/s.

4. Benchmarks

Scope: The scope of the NoCrash and the LeaderBoard benchmark are illustrated in Table 1. As the latest benchmark on CARLA, the LeaderBoard benchmark considers more traffic scenarios and longer routes in six different maps. In this paper we use the publicly available training and testing routes of the LeaderBoard.

Weather: Following the NoCrash benchmark, we use *ClearNoon*, *WetNoon*, *HardRainNoon* and *ClearSunset* as the training weather types, whereas new weather types are *SoftRainSunset* and *WetSunset*. To save computational resources, only two out of the four training weather types are evaluated, they are *WetNoon* and *ClearSunset*.

Background Traffic: The number of vehicles and pedestrians spawned in each map of different benchmarks are listed in Table 2. Vehicles and pedestrians are spawned randomly from the complete blueprint library of CARLA 0.9.11. This stands in contrast to several previous works where for example two-wheeled vehicles are disabled.

5. Additional Experimental Results

To verify IL agents trained using the feature loss indeed embed camera images to the latent space of Roach, we report the feature loss at test time in Fig. 2. In the first row of Fig. 2, the IL agent trained without feature loss, $\mathcal{L}_{\text{K}}$, learns a latent space independent of the one of Roach. Hence, the test feature loss is effectively noise that is invariant to the test condition. In the second row, $\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}(\text{c})$ is trained with the feature loss. The test feature loss of this agent is much smaller (less than 1) and increases as expected during the generalization tests.

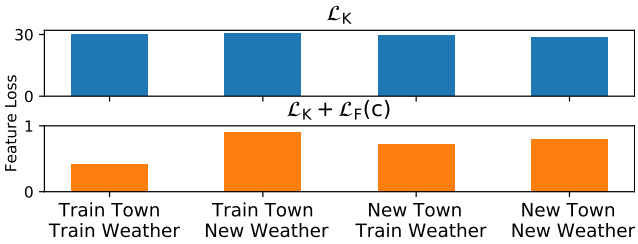


Figure 2: **Feature loss w.r.t. Roach** on one of the NoCrash-dense route. The y-axis of both charts have different scale.

Map	# Routes	Total Km	# Traffic lights	# Stop signs
NoCrash Train				
Town01	25	17.4	110	0
NoCrash Test				
Town02	25	8.9	94	0
LeaderBoard Train				
Town01	10	7.9	47	0
Town03	20	30.7	140	63
Town04	10	24.1	72	13
Town06	10	19.5	58	1
LeaderBoard Test				
Town02	6	5.5	54	0
Town04	10	24.1	72	13
Town05	10	12.4	82	29

Table 1: **Scope of the Nocrash benchmark and the LeaderBoard benchmark.** Total kilometers, number of traffic lights and stop signs are measured using Roach.

To complete Fig. 5 of the main paper, driving scores of experts and IL agents at each DAGGER iterations are in

- Fig. 3: NoCrash-busy.
- Fig. 4: LeaderBoard-busy.

To complete Table 3 of the main paper, detailed driving performance and infraction analysis of our experts and IL agents (5th DAGGER iteration) are listed in

- Table 7: NoCrash-busy, train town & train weather.
- Table 8: NoCrash-busy, train town & new weather.
- Table 9: NoCrash-busy, new town & train weather.
- Table 10: NoCrash-busy, new town & new weather.
- Table 11: LeaderBoard, train town & train weather.
- Table 12: LeaderBoard, train town & new weather.
- Table 13: LeaderBoard, new town & train weather.
- Table 14: LeaderBoard, new town & new weather.

Map	# Vehicles	# Pedestrians
NoCrash dense		
Town01	100	250
Town02	70	150
NoCrash busy		
Town01	120	120
Town02	70	70
LeaderBoard busy		
Town01	120	120
Town02	70	70
Town03	70	70
Town04	150	80
Town05	120	120
Town06	120	80

Table 2: **Background traffic settings for different benchmarks.**

Layer Type	Filters	Size	Strides	Activation
Image Encoder				
Conv2d	8	5x5	2	ReLU
Conv2d	16	5x5	2	ReLU
Conv2d	32	5x5	2	ReLU
Conv2d	64	3x3	2	ReLU
Conv2d	128	3x3	2	ReLU
Conv2d	256	3x3	1	-
Flatten				
Measurement Encoder				
Dense	256			ReLU
Dense	256			ReLU
FC Layers after Concatenation				
Dense	512			ReLU
Dense	256			ReLU
Action Head				
Dense (shared)	256			ReLU
Dense (shared)	256			ReLU
Dense (for α)	2			Softplus
Dense (for β)	2			Softplus
Value Head				
Dense	256			ReLU
Dense	256			ReLU
Dense	1			-

Table 3: **The network architecture used for Roach.** Around 1.53M trainable parameters.

Layer Type	Filters	Activation	Dropout
Image Encoder			
ResNet-34			
Measurement Encoder			
Dense	128	ReLU	
Dense	128	ReLU	
FC Layers after concatenation			
Dense	512	ReLU	
Dense	512	ReLU	
Dense	256	ReLU	
Speed Head			
Dense	256	ReLU	
Dense	256	ReLU	0.5
Dense	1		
Value Head			
Dense	256	ReLU	
Dense	256	ReLU	0.5
Dense	1		
Deterministic Action Head			
Dense	256	ReLU	
Dense	256	ReLU	0.5
Dense	2		
Distributional Action Head			
Dense (shared)	256	ReLU	
Dense (shared)	256	ReLU	0.5
Dense (for α)	2	Softplus	
Dense (for β)	2	Softplus	

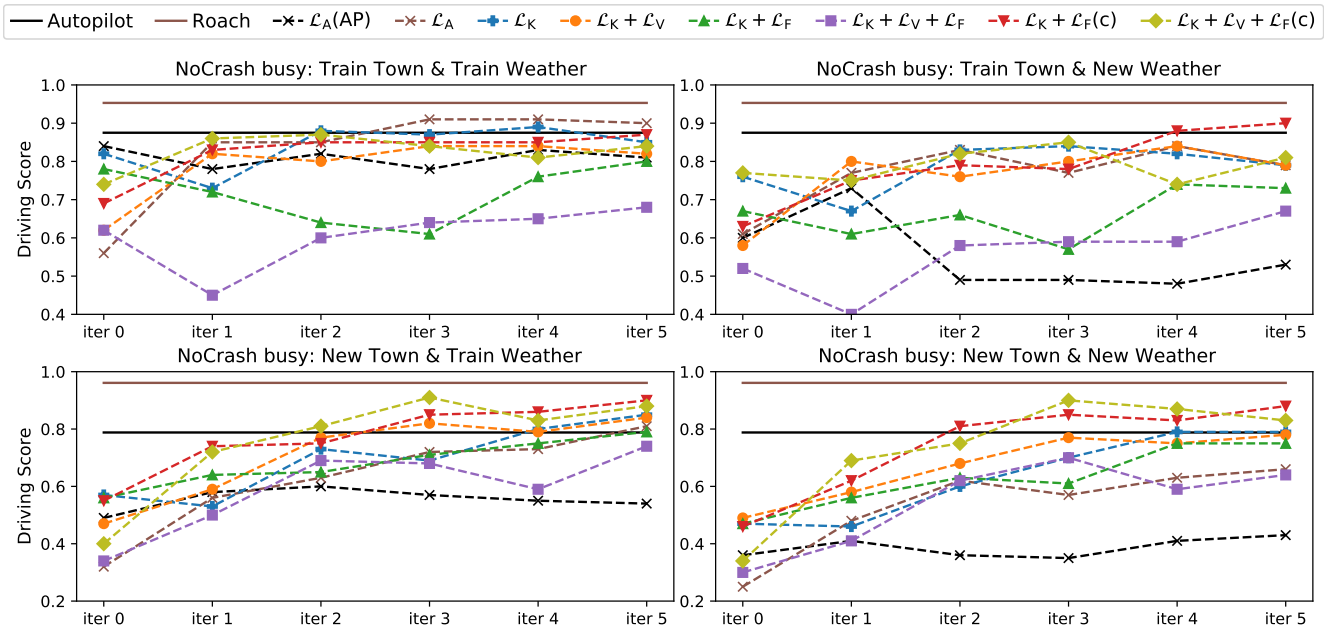
Table 4: **The network architecture used for our IL agent.** Around 23.4M trainable parameters.

Notation	Description	Value
	BEV Representation	
W	Width	192 px
H	Height	192 px
C	Number of channels	15
K	Size of the temporal sequence	4
	Timestamps of images in the temporal sequence	{-1.5, -1, -0.5, 0} sec
D	Distance from the ego-vehicle to the bottom	40 px
	Pixels per meter	5 px/m
	Minimum width/height of rendered bounding boxes	8 px
	Scale factor for bounding box size of pedestrians	2
	Rollout	
	Buffer size for six environments	12288 frames
	Value bootstrap for the last non-terminal sample	True
	Synchronized	True
	Reset at the beginning of a new phase	False
	Weather	dynamic
	Range of vehicle/pedestrian number in Town 1	$[0, 150]/[0, 300]$
	Range of vehicle/pedestrian number in Town 2	$[0, 100]/[0, 200]$
	Range of vehicle/pedestrian number in Town 3	$[0, 120]/[0, 120]$
	Range of vehicle/pedestrian number in Town 4	$[0, 160]/[0, 160]$
	Range of vehicle/pedestrian number in Town 5	$[0, 160]/[0, 160]$
	Range of vehicle/pedestrian number in Town 6	$[0, 160]/[0, 160]$
	Update	
	Number of epochs	20
λ_{ent}	Weight for the entropy loss	0.01
λ_{exp}	Weight for the exploration loss	0.05
	Weight for value loss	0.5
	γ for GAE	0.99
	λ for GAE	0.9
	Clipping range for PPO-clip	0.2
	Max norm for gradient clipping	0.5
	Batch size	256
	Initial learning rate	1e-5
	KL-divergence threshold for learning rate schedule	0.15
	Patience for learning rate schedule	8
	Factor for learning rate schedule	0.5

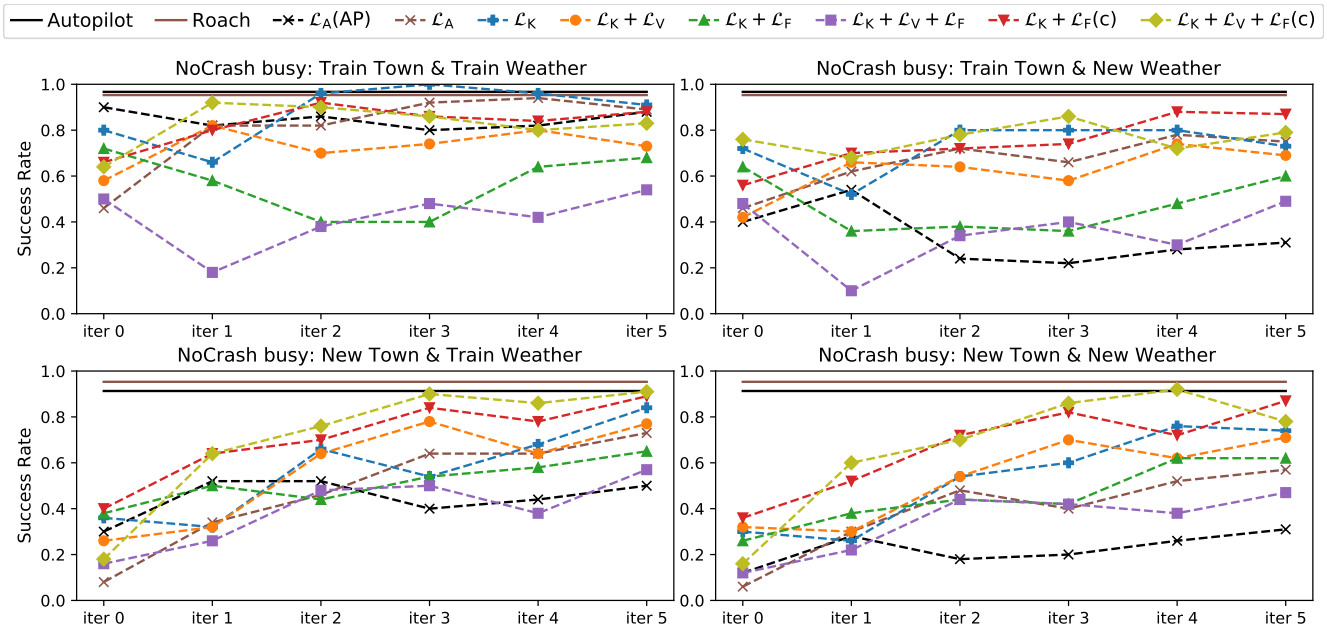
Table 5: **The hyper-parameter values used for Roach.**

Description	Value
Inputs	
Camera type	RGB
Camera image width	900 px
Camera image height	256 px
Camera location $[x, y, z]$ relative to the ego-vehicle	$[-1.5, 0, 2]$
Camera rotation $[roll, pitch, yaw]$ relative to the ego-vehicle	$[0, 0, 0]$
Camera horizontal FOV	$100°$
Mean for image normalization	$[0.485, 0.456, 0.406]$
Standard deviation for image normalization	$[0.229, 0.224, 0.225]$
Speed measurement	Forward speed in m/s
Normalization factor for speed	12
Data Collection	
Episode length	300 sec
Triangular perturbation for off-policy data	20%
Number of episodes (NoCrash, off-policy)	80
Number of episodes (LeaderBoard, off-policy)	160
Number of episodes (NoCrash, on-policy, Autopilot)	80
Number of episodes (LeaderBoard, on-policy, Autopilot)	160
Number of episodes (NoCrash, on-policy, Roach)	40
Number of episodes (LeaderBoard, on-policy, Roach)	80
DA-RB critical state sampling criterion	difference in acceleration
DA-RB critical state sampling threshold	0.2
Weather	Same as NoCrash train weathers
Range of vehicle/pedestrian number in NoCrash train town 1	$[0, 150]/[0, 200]$
Range of vehicle/pedestrian number in LeaderBoard train town 1	$[80, 160]/[80, 160]$
Range of vehicle/pedestrian number in LeaderBoard train town 3	$[40, 100]/[40, 100]$
Range of vehicle/pedestrian number in LeaderBoard train town 4	$[100, 200]/[40, 120]$
Range of vehicle/pedestrian number in LeaderBoard train town 6	$[80, 160]/[40, 120]$
Training	
Number of epochs at each DAGGER iteration	25
λ_{S}, weight for the speed regularization	0.05
λ_{V}, weight for the value loss, if applied	0.05
λ_{F}, weight for the feature loss, if applied	0.001
Batch size	48
Initial learning rate	0.0002
Patience for reduce-on-plateau learning rate schedule	5
Factor for learning rate schedule	0.5
Pre-trained distributional action head	True
Pre-trained value head	True
Image augmentation	True

Table 6: **The hyper-parameter values used for our IL agent.**

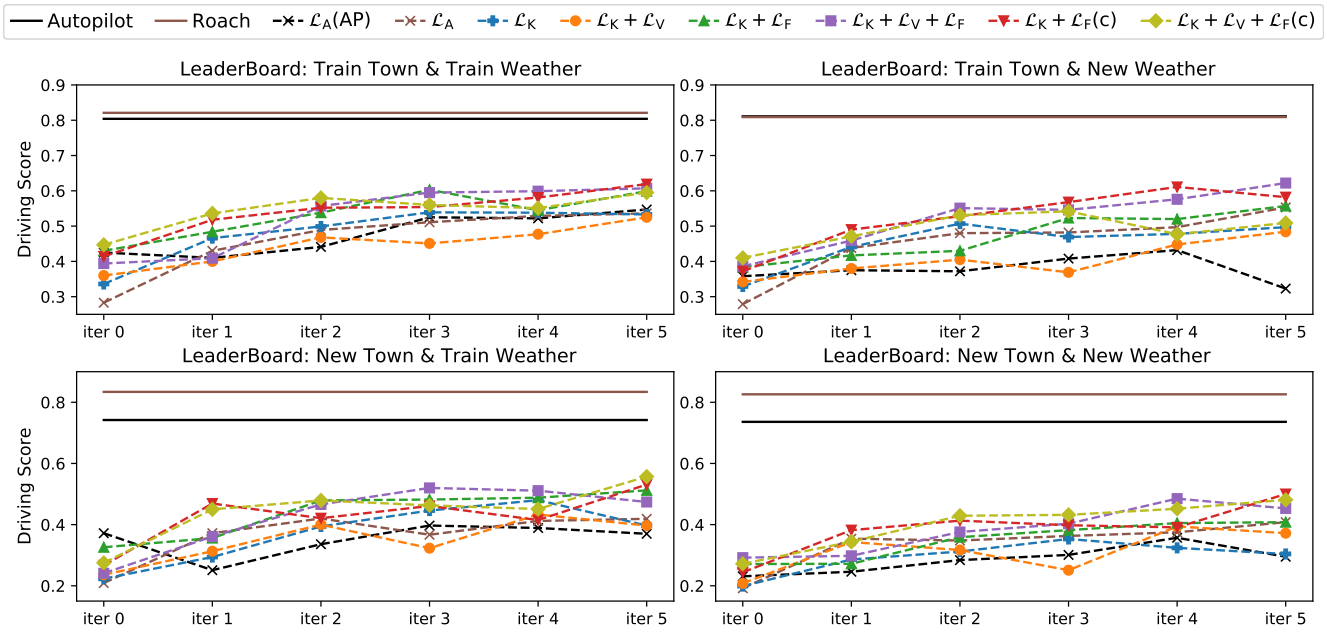


(a) **Driving Score**

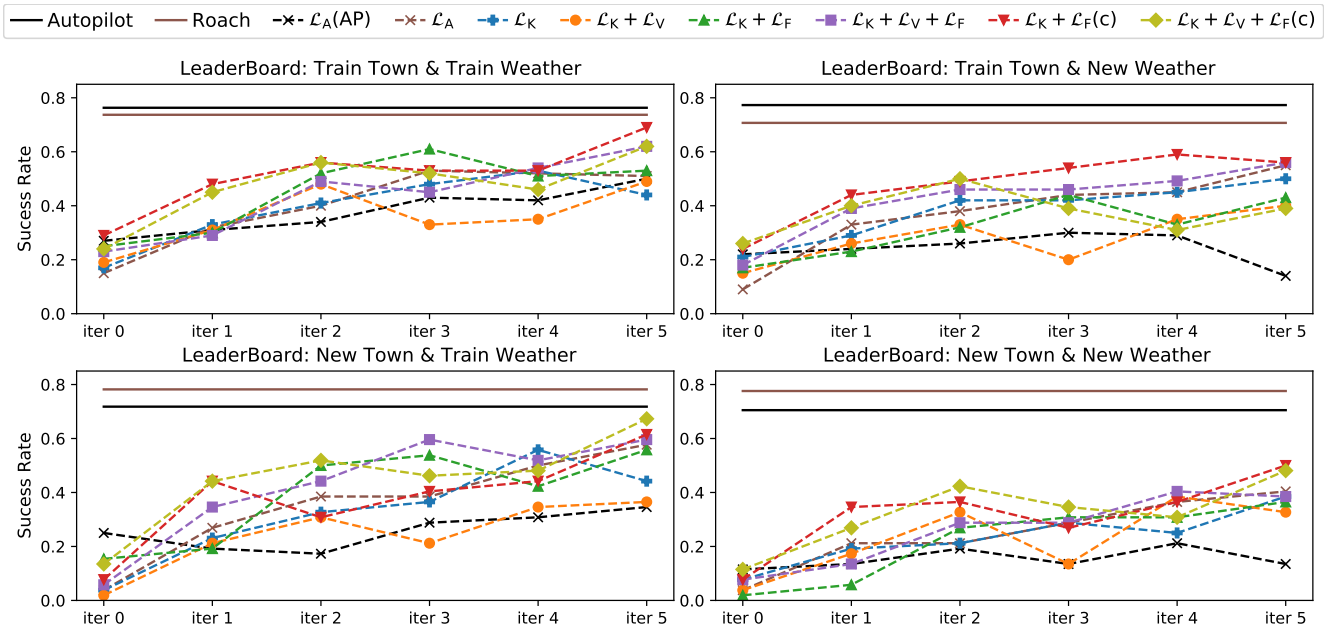


(b) **Success Rate**

Figure 3: **Driving performance of experts and IL agents on the NoCrash-busy benchmark.** All IL agents (dashed lines) are supervised by Roach except for $\mathcal{L}_A$(AP), which is supervised by the CARLA Autopilot. For IL agents at the 5th iteration and all experts, results are reported as the mean over 3 evaluation seeds. Others agents are evaluated only once.



(a) **Driving Score**



(b) **Success Rate**

Figure 4: **Driving performance of experts and IL agents on the LeaderBoard-busy benchmark.** All IL agents (dashed lines) are supervised by Roach except for $\mathcal{L}_A$(AP), which is supervised by the CARLA Autopilot. For all experts, results are reported as the mean over 3 evaluation seeds. Results of IL agents are evaluated only once.

	Success rate	Driving score	Route compl.	Infrac. penalty	Collision others	Collision pedestrian	Collision vehicle	Red light infraction	Agent blocked
iter 5	%, ↑	%, ↑	%, ↑	%, ↑	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓
$\mathcal{L}_{\text{A}}$(AP)	88 ± 2	81 ± 2	94 ± 2	86 ± 1	$\mathbf{0} \pm 0$	$\mathbf{0} \pm 0$	0.08 ± 0.11	1.02 ± 0.33	1 ± 0.28
$\mathcal{L}_{\text{A}}$	89 ± 5	$\mathbf{90} \pm 2$	$\mathbf{99} \pm 1$	90 ± 1	0.06 ± 0.04	0.05 ± 0.02	0.06 ± 0.04	0.29 ± 0.03	$\mathbf{0.05} \pm 0.06$
$\mathcal{L}_{\text{K}}$	$\mathbf{91} \pm 10$	85 ± 6	99 ± 2	85 ± 5	0.1 ± 0.18	0.03 ± 0.04	0.1 ± 0.11	0.58 ± 0.07	0.07 ± 0.12
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}}$	73 ± 4	82 ± 3	91 ± 2	$\mathbf{91} \pm 2$	0.07 ± 0.07	0.02 ± 0.02	0.18 ± 0.12	$\mathbf{0.27} \pm 0.06$	0.6 ± 0.2
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$	68 ± 11	80 ± 6	89 ± 3	89 ± 4	0.15 ± 0.03	0.02 ± 0.01	$\mathbf{0.05} \pm 0.06$	0.41 ± 0.13	0.12 ± 0.02
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$	54 ± 2	68 ± 3	80 ± 2	87 ± 3	0.22 ± 0.34	0.06 ± 0.05	0.08 ± 0.05	0.53 ± 0.08	0.91 ± 0.32
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$(c)	88 ± 2	87 ± 2	98 ± 1	88 ± 2	0.05 ± 0.08	0.07 ± 0.02	0.1 ± 0.07	0.41 ± 0.05	0.33 ± 0.49
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$(c)	83 ± 1	84 ± 2	95 ± 1	89 ± 3	$\mathbf{0} \pm 0$	0.04 ± 0.03	0.06 ± 0.06	0.5 ± 0.16	0.06 ± 0.06
Roach	95 ± 5	95 ± 1	100 ± 0	95 ± 1	0 ± 0	0.04 ± 0.04	0.03 ± 0.04	0.13 ± 0.11	0 ± 0
Autopilot	97 ± 2	87 ± 4	99 ± 2	88 ± 3	0 ± 0	0 ± 0	0.33 ± 0.55	0.89 ± 0.54	0.35 ± 0.58

Table 7: **Performance and infraction analysis on NoCrash-busy, train town & train weather.** Mean and std. over 3 seeds.

	Success rate	Driving score	Route compl.	Infrac. penalty	Collision others	Collision pedestrian	Collision vehicle	Red light infraction	Agent blocked
iter 5	%, ↑	%, ↑	%, ↑	%, ↑	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓
$\mathcal{L}_{\text{A}}$(AP)	31 ± 3	53 ± 2	61 ± 1	87 ± 1	$\mathbf{0} \pm 0$	$\mathbf{0} \pm 0$	0.35 ± 0.23	1.31 ± 0.36	5.75 ± 0.11
$\mathcal{L}_{\text{A}}$	75 ± 4	79 ± 5	92 ± 2	87 ± 4	0.13 ± 0.18	0.03 ± 0	0.06 ± 0.01	0.69 ± 0.19	0.79 ± 0.32
$\mathcal{L}_{\text{K}}$	73 ± 5	79 ± 5	91 ± 3	87 ± 3	0.02 ± 0.04	$\mathbf{0} \pm 0$	0.24 ± 0.37	0.6 ± 0.18	0.95 ± 0.45
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}}$	69 ± 4	79 ± 3	91 ± 1	87 ± 3	0.03 ± 0.05	0.04 ± 0.03	0.14 ± 0.07	0.5 ± 0.1	0.7 ± 0.05
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$	60 ± 5	73 ± 2	80 ± 3	92 ± 1	0.05 ± 0.08	0.1 ± 0.16	0.09 ± 0.05	0.38 ± 0.03	$\mathbf{0.02} \pm 0.03$
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$	49 ± 8	67 ± 4	75 ± 4	90 ± 1	0.07 ± 0.13	0.03 ± 0.05	0.86 ± 1.41	0.88 ± 0.61	0.73 ± 0.17
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$(c)	$\mathbf{87} \pm 5$	$\mathbf{90} \pm 2$	$\mathbf{97} \pm 2$	$\mathbf{93} \pm 1$	$\mathbf{0} \pm 0$	0.01 ± 0.03	0.03 ± 0.06	$\mathbf{0.37} \pm 0.03$	0.23 ± 0.13
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$(c)	79 ± 3	81 ± 0	92 ± 1	89 ± 1	$\mathbf{0} \pm 0$	0.01 ± 0.01	$\mathbf{0.02} \pm 0.02$	0.57 ± 0.06	0.39 ± 0.17
Roach	95 ± 5	95 ± 1	100 ± 0	95 ± 1	0 ± 0	0.04 ± 0.04	0.03 ± 0.04	0.13 ± 0.11	0 ± 0
Autopilot	97 ± 2	87 ± 4	99 ± 2	88 ± 3	0 ± 0	0 ± 0	0.33 ± 0.55	0.89 ± 0.54	0.35 ± 0.58

Table 8: **Performance and infraction analysis on NoCrash-busy, train town & new weather.** Mean and std. over 3 seeds.

	Success rate	Driving score	Route compl.	Infrac. penalty	Collision others	Collision pedestrian	Collision vehicle	Red light infraction	Agent blocked
iter 5	%, ↑	%, ↑	%, ↑	%, ↑	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓
$\mathcal{L}_{\text{A}}$(AP)	50 ± 5	54 ± 1	79 ± 3	72 ± 3	0.88 ± 0.86	$\mathbf{0} \pm 0$	$\mathbf{0.08} \pm 0.06$	3.24 ± 0.35	3.76 ± 0.8
$\mathcal{L}_{\text{A}}$	73 ± 4	81 ± 4	94 ± 4	85 ± 2	1.03 ± 1.09	0.09 ± 0.05	0.72 ± 0.8	0.79 ± 0.12	1.24 ± 0.88
$\mathcal{L}_{\text{K}}$	84 ± 7	85 ± 4	97 ± 1	88 ± 4	0.25 ± 0.13	0.02 ± 0.03	0.3 ± 0.31	0.74 ± 0.18	0.37 ± 0.04
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}}$	77 ± 10	84 ± 5	97 ± 3	86 ± 3	0.25 ± 0.28	0.02 ± 0.03	0.49 ± 0.13	0.73 ± 0.18	0.19 ± 0.24
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$	65 ± 2	79 ± 2	88 ± 1	90 ± 3	0.31 ± 0.47	0.07 ± 0.07	0.37 ± 0.16	0.6 ± 0.19	0.3 ± 0.45
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$	57 ± 4	74 ± 4	82 ± 1	90 ± 4	0.96 ± 0.2	0.04 ± 0.05	0.22 ± 0.16	$\mathbf{0.43} \pm 0.21$	0.93 ± 0.23
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$(c)	89 ± 5	$\mathbf{90} \pm 3$	$\mathbf{100} \pm 1$	$\mathbf{90} \pm 2$	$\mathbf{0.02} \pm 0.03$	0.08 ± 0.07	0.23 ± 0.11	0.59 ± 0.12	$\mathbf{0.04} \pm 0.08$
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$(c)	$\mathbf{91} \pm 5$	88 ± 4	98 ± 1	89 ± 3	0.06 ± 0.06	0.01 ± 0.03	0.19 ± 0.08	0.78 ± 0.25	0.06 ± 0.06
Roach	95 ± 2	96 ± 3	100 ± 0	96 ± 3	0 ± 0	0.11 ± 0.07	0.04 ± 0.05	0.16 ± 0.2	0 ± 0
Autopilot	91 ± 1	79 ± 2	98 ± 1	80 ± 2	0 ± 0	0 ± 0	0.18 ± 0.08	1.93 ± 0.23	0.18 ± 0.08

Table 9: **Performance and infraction analysis on NoCrash-busy, new town & train weather.** Mean and std. over 3 seeds.

	Success rate	Driving score	Route compl.	Infrac. penalty	Collision others	Collision pedestrian	Collision vehicle	Red light infraction	Agent blocked
iter 5	%, ↑	%, ↑	%, ↑	%, ↑	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓
$\mathcal{L}_{\mathrm{A}}$(AP)	31 ± 7	43 ± 2	62 ± 6	77 ± 4	0.54 ± 0.53	$\mathbf{0} \pm 0$	0.63 ± 0.50	3.33 ± 0.58	19.4 ± 14.4
$\mathcal{L}_{\mathrm{A}}$	57 ± 7	66 ± 3	84 ± 3	76 ± 1	2.07 ± 1.37	$\mathbf{0} \pm 0$	1.36 ± 1.10	1.4 ± 0.2	2.82 ± 1.45
$\mathcal{L}_{\mathrm{K}}$	74 ± 3	79 ± 0	91 ± 2	86 ± 1	0.50 ± 0.25	$\mathbf{0} \pm 0$	0.53 ± 0.18	0.68 ± 0.08	3.39 ± 0.20
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}}$	71 ± 9	78 ± 3	91 ± 1	85 ± 3	0.55 ± 0.22	0.11 ± 0.06	0.34 ± 0.31	0.72 ± 0.09	1.14 ± 0.10
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{F}}$	62 ± 2	75 ± 1	85 ± 0	87 ± 2	0.79 ± 0.61	0.03 ± 0.05	0.73 ± 0.16	0.63 ± 0.02	2.04 ± 1.33
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}} + \mathcal{L}_{\mathrm{F}}$	47 ± 9	64 ± 6	72 ± 5	89 ± 3	0.9 ± 0.73	0.03 ± 0.06	0.38 ± 0.26	0.79 ± 0.42	1.29 ± 0.9
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{F}}$(c)	$\mathbf{87} \pm 5$	$\mathbf{88} \pm 3$	$\mathbf{96} \pm 0$	$\mathbf{91} \pm 3$	$\mathbf{0.08} \pm 0.04$	0.01 ± 0.02	0.23 ± 0.08	$\mathbf{0.61} \pm 0.23$	0.84 ± 0.04
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}} + \mathcal{L}_{\mathrm{F}}$(c)	78 ± 3	83 ± 1	94 ± 2	89 ± 2	0.21 ± 0.14	$\mathbf{0} \pm 0$	$\mathbf{0.16} \pm 0.05$	0.79 ± 0.15	$\mathbf{0.46} \pm 0.13$
Roach	95 ± 2	96 ± 3	100 ± 0	96 ± 3	0 ± 0	0.11 ± 0.07	0.04 ± 0.05	0.16 ± 0.20	0 ± 0
Autopilot	91 ± 1	79 ± 2	98 ± 1	80 ± 2	0 ± 0	0 ± 0	0.18 ± 0.08	1.93 ± 0.23	0.18 ± 0.08

Table 10: **Performance and infraction analysis on NoCrash-busy, new town & new weather.** Mean and std. over 3 seeds.

	Success rate	Driving score	Route compl.	Infrac. penalty	Collision others	Collision pedestrian	Collision vehicle	Red light infraction	Stop Sign infraction	Agent blocked
iter 5	%, ↑	%, ↑	%, ↑	%, ↑	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓
$\mathcal{L}_{\mathrm{A}}$(AP)	50	55	82	68	0.24	**0.01**	0.38	0.53	**0.22**	1.39
$\mathcal{L}_{\mathrm{A}}$	51	54	87	60	0.46	0.19	0.30	0.50	0.39	0.48
$\mathcal{L}_{\mathrm{K}}$	44	53	86	63	0.14	0.07	0.35	0.42	0.38	0.77
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}}$	49	53	81	66	0.39	0.04	0.30	0.36	0.40	1.35
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{F}}$	53	60	85	**71**	0.11	0.10	0.20	**0.25**	0.32	0.47
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}} + \mathcal{L}_{\mathrm{F}}$	62	61	94	65	**0.01**	0.05	0.30	0.37	0.42	0.47
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{F}}$(c)	**69**	**62**	94	66	0.05	0.04	**0.15**	0.35	0.59	**0.40**
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}} + \mathcal{L}_{\mathrm{F}}$(c)	62	59	**95**	63	0.04	0.41	0.21	0.33	0.50	0.45
Roach	74 ± 1	82 ± 2	95 ± 1	86 ± 2	0.03 ± 0.02	0.04 ± 0.03	0.12 ± 0.04	0.13 ± 0.05	0 ± 0.01	0.13 ± 0.04
Autopilot	76 ± 1	80 ± 1	96 ± 1	84 ± 2	0 ± 0	0 ± 0	0.16 ± 0.05	0.3 ± 0.05	0 ± 0.01	0.16 ± 0.07

Table 11: **Performance and infraction analysis on the LeaderBoard, train town & train weather.** Mean and std. over 3 seeds.

	Success rate	Driving score	Route compl.	Infrac. penalty	Collision others	Collision pedestrian	Collision vehicle	Red light infraction	Stop Sign infraction	Agent blocked
iter 5	%, ↑	%, ↑	%, ↑	%, ↑	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓
$\mathcal{L}_{\mathrm{A}}$(AP)	14	32	47	**79**	0.23	**0.00**	0.31	0.55	0.32	31.79
$\mathcal{L}_{\mathrm{A}}$	55	55	87	64	0.14	0.03	0.26	0.37	0.47	0.43
$\mathcal{L}_{\mathrm{K}}$	50	50	87	58	0.08	0.06	0.42	0.57	0.62	0.61
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}}$	40	48	79	64	0.13	0.02	0.37	0.48	0.45	0.80
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{F}}$	43	56	82	70	0.11	0.03	0.20	0.34	**0.31**	0.66
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}} + \mathcal{L}_{\mathrm{F}}$	**56**	**62**	**91**	69	**0.06**	0.05	**0.15**	**0.29**	0.39	**0.31**
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{F}}$(c)	**56**	58	90	66	0.07	0.05	0.19	0.36	0.60	0.36
$\mathcal{L}_{\mathrm{K}} + \mathcal{L}_{\mathrm{V}} + \mathcal{L}_{\mathrm{F}}$(c)	39	51	88	59	0.10	0.03	0.29	0.38	0.54	0.47
Roach	71 ± 2	81 ± 1	95 ± 1	85 ± 0	0.02 ± 0.02	0.04 ± 0.03	0.14 ± 0.02	0.12 ± 0.04	0 ± 0.01	0.14 ± 0.07
Autopilot	77 ± 2	81 ± 1	96 ± 1	85 ± 2	0 ± 0	0 ± 0	0.16 ± 0.04	0.28 ± 0.06	0 ± 0.01	0.22 ± 0.13

Table 12: **Performance and infraction analysis on the LeaderBoard, train town & new weather.** Mean and std. over 3 seeds.

	Success rate	Driving score	Route compl.	Infrac. penalty	Collision others	Collision pedestrian	Collision vehicle	Red light infraction	Stop Sign infraction	Agent blocked
iter 5	%, ↑	%, ↑	%, ↑	%, ↑	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓
$\mathcal{L}_{\text{A}}$(AP)	35	37	75	55	0.17	**0.00**	1.52	1.00	0.50	3.64
$\mathcal{L}_{\text{A}}$	58	42	92	46	0.17	0.04	0.42	0.82	0.75	0.29
$\mathcal{L}_{\text{K}}$	44	40	91	44	0.91	0.04	0.36	1.21	0.75	0.94
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}}$	37	40	76	**58**	0.13	0.05	0.45	0.80	**0.40**	0.33
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$	56	51	91	56	0.22	0.07	0.34	0.45	0.61	0.16
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$	60	47	**95**	50	**0.00**	0.09	0.81	0.64	0.74	1.29
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$(c)	62	53	94	56	0.00	0.04	1.22	0.71	0.70	1.04
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$(c)	**67**	**56**	**95**	**58**	0.03	0.05	**0.31**	**0.38**	0.72	**0.11**
Roach	78 ± 4	83 ± 2	97 ± 1	86 ± 2	0 ± 0	0.03 ± 0.02	0.13 ± 0.1	0.16 ± 0.03	0 ± 0	0.09 ± 0.04
Autopilot	72 ± 13	74 ± 5	95 ± 2	78 ± 3	0 ± 0	0 ± 0	0.14 ± 0.07	0.57 ± 0.13	0 ± 0	0.18 ± 0.14

Table 13: **Performance and infraction analysis on the LeaderBoard, new town & train weather.** Mean and std. over 3 seeds.

	Success rate	Driving score	Route compl.	Infrac. penalty	Collision others	Collision pedestrian	Collision vehicle	Red light infraction	Stop Sign infraction	Agent blocked
iter 5	%, ↑	%, ↑	%, ↑	%, ↑	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓	#/Km, ↓
$\mathcal{L}_{\text{A}}$(AP)	14	30	42	**80**	0.06	0.05	1.11	0.63	**0.49**	28.27
$\mathcal{L}_{\text{A}}$	40	41	**90**	45	0.36	0.10	0.61	0.88	0.67	**0.35**
$\mathcal{L}_{\text{K}}$	39	30	86	39	0.17	0.03	0.42	1.31	0.81	0.51
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}}$	33	37	78	53	0.09	0.06	0.47	0.97	0.54	0.40
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$	37	41	81	54	0.29	0.03	0.79	0.61	0.68	1.23
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$	39	45	85	56	0.02	**0.01**	1.54	0.73	0.64	2.30
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{F}}$(c)	**50**	**50**	86	60	**0.01**	0.02	0.48	**0.60**	0.63	2.64
$\mathcal{L}_{\text{K}} + \mathcal{L}_{\text{V}} + \mathcal{L}_{\text{F}}$(c)	48	48	**90**	56	0.02	0.04	**0.18**	**0.60**	0.81	0.47
Roach	78 ± 4	83 ± 2	97 ± 1	85 ± 2	0 ± 0	0.04 ± 0.02	0.13 ± 0.1	0.18 ± 0.06	0 ± 0	0.09 ± 0.04
Autopilot	71 ± 11	74 ± 4	95 ± 1	78 ± 3	0 ± 0	0 ± 0	0.14 ± 0.07	0.58 ± 0.12	0 ± 0	0.2 ± 0.12

Table 14: **Performance and infraction analysis on the LeaderBoard, new town & new weather.** Mean and std. over 3 seeds.

References

[1] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016. 1

[2] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9329–9338, 2019. 3

[3] Aditya Prakash, Aseem Behl, Eshed Ohn-Bar, Kashyap Chitta, and Andreas Geiger. Exploring data aggregation in policy learning for vision-based urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11763–11773, 2020. 3

[4] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. Is deep reinforcement learning really superhuman on atari? In *Deep Reinforcement Learning Workshop of the Conference on Neural Information Processing Systems*, 2019. 2