Aveiro, October 04 2024

End-to-End Reinforcement Learning for Autonomous Driving in Urban Environments

Thesis Defense

Ph.D. candidate: **Daniel Coelho**

Supervisor: **Miguel Oliveira**

Co-supervisor: **Vítor Santos**

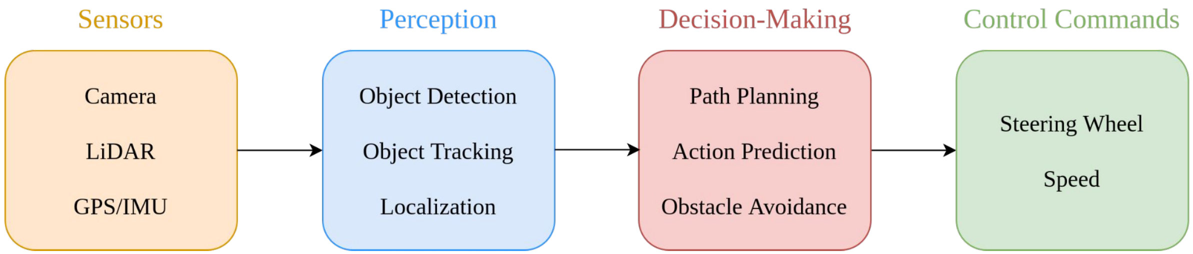
Outline

1.

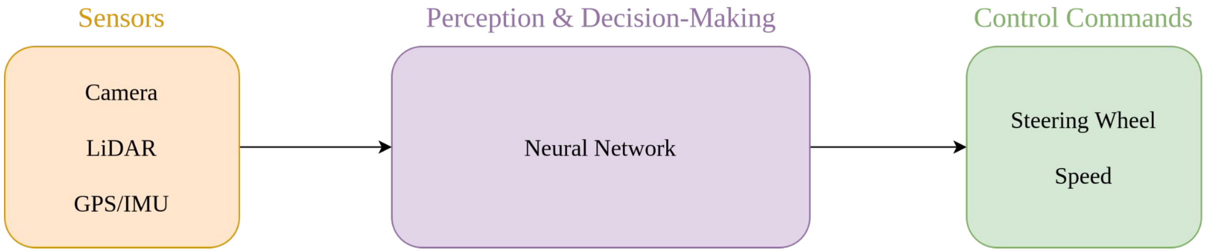
Introduction

Coelho, Daniel, and Miguel Oliveira. "A review of end-to-end autonomous driving in urban environments." **IEEE Access (2022)**: 75296-75311, doi: 10.1109/ACCESS.2022.3192019.

Autonomous Driving



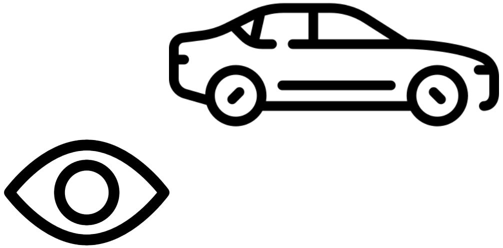
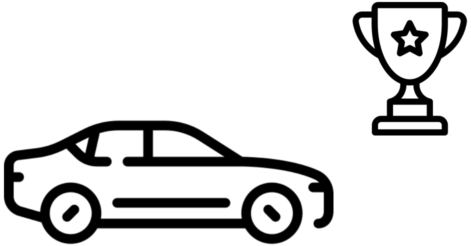
Modular [1]



End-to-End [2]

[1] C. Urmson et al., "Autonomous driving in urban environments: Boss and the urban challenge," J. Field Robot., vol. 25, pp. 425–466, 2009.
[2] M. Bojarski et al., "End to end learning for self-driving cars," 2016, arXiv:160407316.

End-to-End Autonomous Driving

Imitation Learning (IL) [2]	**Reinforcement Learning (RL)** [3]
	

[2] K. Chitta et al., “Transfuser: Imitation with transformer-based sensor fusion for autonomous driving”, in IEEE TPAMI, 2022, 45(11), 12878-12895.
[3] A. Kendall et al., ‘‘Learning to drive in a day,’’ in Proc. Int. Conf. Robot. Autom. (ICRA), 2019, pp. 8248–8254.

Simulation Framework

- **Real-world** research in AD is **costly**, **risky**, and presents **ethical dilemmas**, making it impractical to rely solely on real-world testing
- **Simulations** provide a **controlled**, **safe**, and **cost-effective** environment for testing diverse driving scenarios that would be difficult or unsafe to replicate in real life
- In this research, the **CARLA simulator** [4], a leading open-source platform, was used for developing, training, and evaluating AD systems

[4] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA:An open urban driving simulator," in Proc. CoRL, 2017, pp. 1–16.

Research Objectives

- Development of End-to-End RL Architectures for AD Systems in Urban Environments
- Integration of Expert Demonstrations in an End-to-End RL Architecture for AD Systems
- Development of a Data-Driven Expert Agent for Improved Driving Simulations

2.

RLAD: Reinforcement Learning from Pixels for Autonomous Driving in Urban Environments

Daniel Coelho, Miguel Oliveira, and Vitor Santos. "RLAD: Reinforcement Learning From Pixels for Autonomous Driving in Urban Environments." **IEEE Transactions on Automation Science and Engineering (2023)**, doi: 10.1109/TASE.2023.3342419.

Motivation

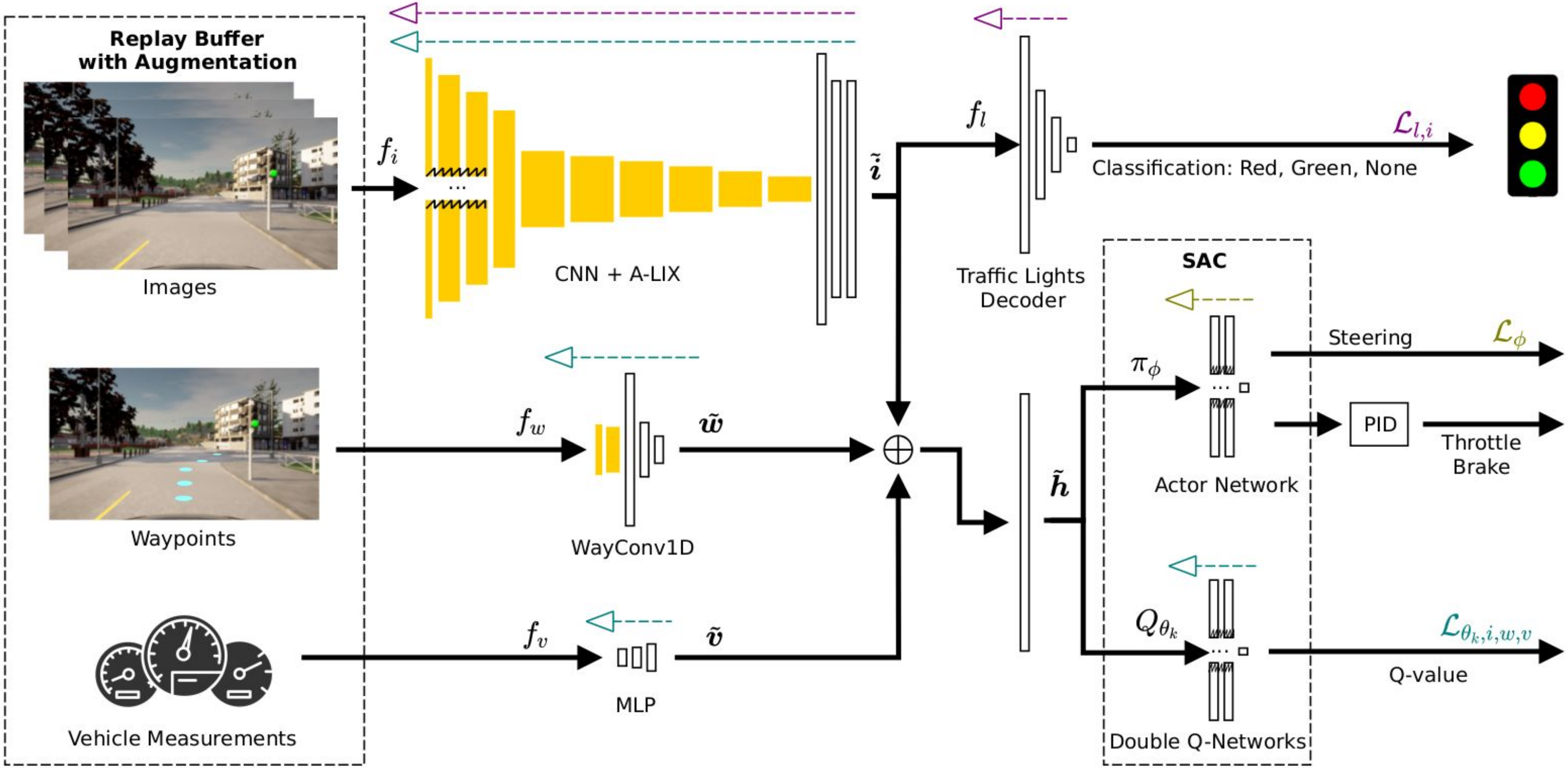
- In urban Autonomous Driving (AD), all methods that use RL train the **encoder** and the **policy separately**
- This is in contrast to Reinforcement Learning from Pixels (RLfP), which trains the **encoder** and the **policy** using the **same objective function**
- By having only one objective function, we ensure that all components are **aligned** with the **downstream task**

Problems of applying RLfP in AD

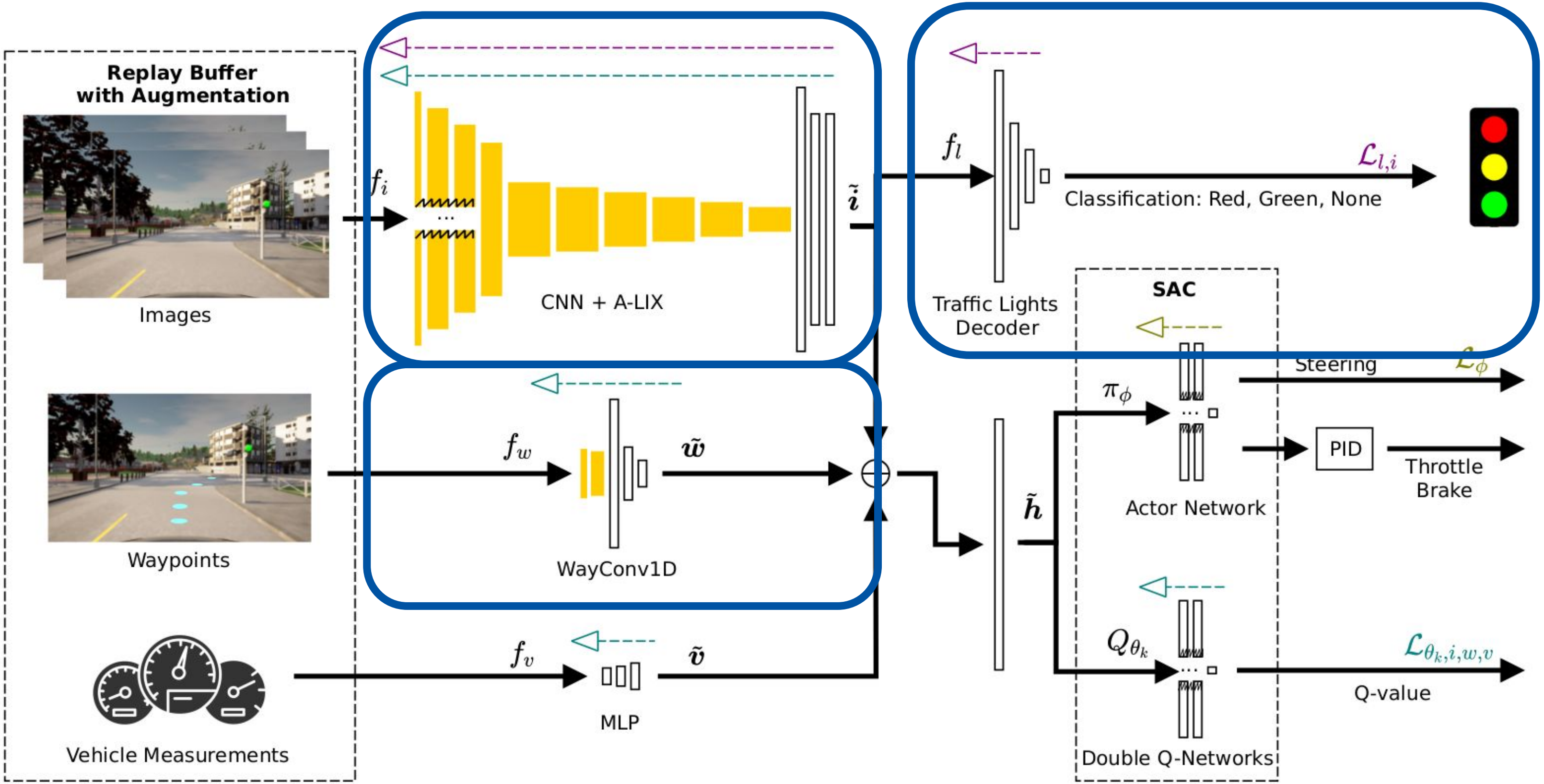
- Sample Inefficiency [5]
- Catastrophic Self-overfitting [6]

[5] D. Yarats et al., "Improving sample efficiency in model-free reinforcement learning from images," in Proc. AAAI Conf. Artif. Intell., May 2021, vol. 35, no. 12,
[6] E. Cetin et al., "Stabilizing off-policy deep reinforcement learning from pixels," in Proc. Int. Conf. Mach. Learn., 2022, pp. 2784–2810.

Architecture



Architecture



Adaptive Local Signal Mixing (A-LIX)

- Technique adapted from [6] , that minimizes the **catastrophic self-overfitting** phenomenon

- A-LIX is applied to features from convolutional layers by **mixing each component with its neighboring components** within the same feature map, using an **exponential weighting** mechanism that reduces the influence of neighbors as the distance increases

- Thus, the feature maps become **spatially consistent**, minimizing the effect of the catastrophic self-overfitting phenomenon.

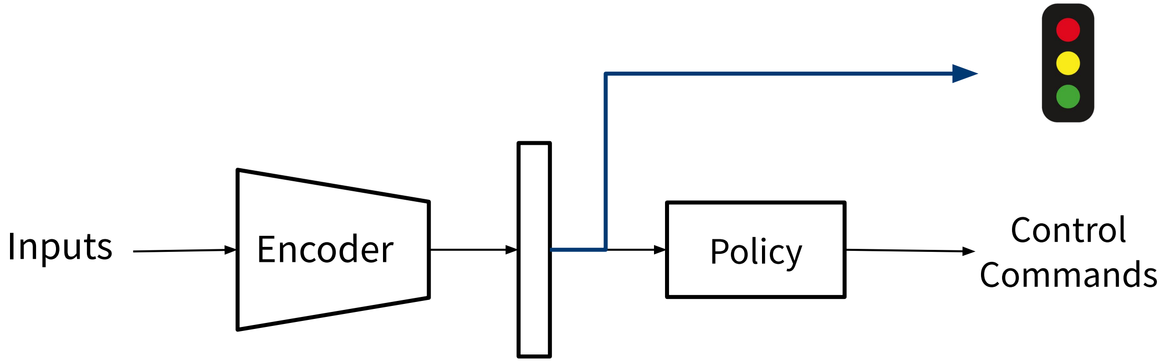
[6] E. Cetin et al., "Stabilizing off-policy deep reinforcement learning from pixels," in Proc. Int. Conf. Mach. Learn., 2022, pp. 2784–2810.

WayConv1D

- WayConv1D is a waypoint encoder that leverages the **2D geometrical structure** of the input by applying **1D convolutions** with a **2×2 kernel** over the 2D coordinates of the next N waypoints
- With WayConv1D the agent learns more efficiently to follow the trajectory without oscillating near the center of the lane.

Traffic Light Decoder

- Auxiliary loss that performs traffic light classification to **strengthen** the **significance of traffic light information** in the **latent representation** of the image



Setup of Experiments: **NoCrash Benchmark**

Town 01

Town 02

NoCrash Benchmark

Empty

Regular

Dense

NoCrash Benchmark

Empty

Regular

Dense

NoCrash Benchmark

Empty

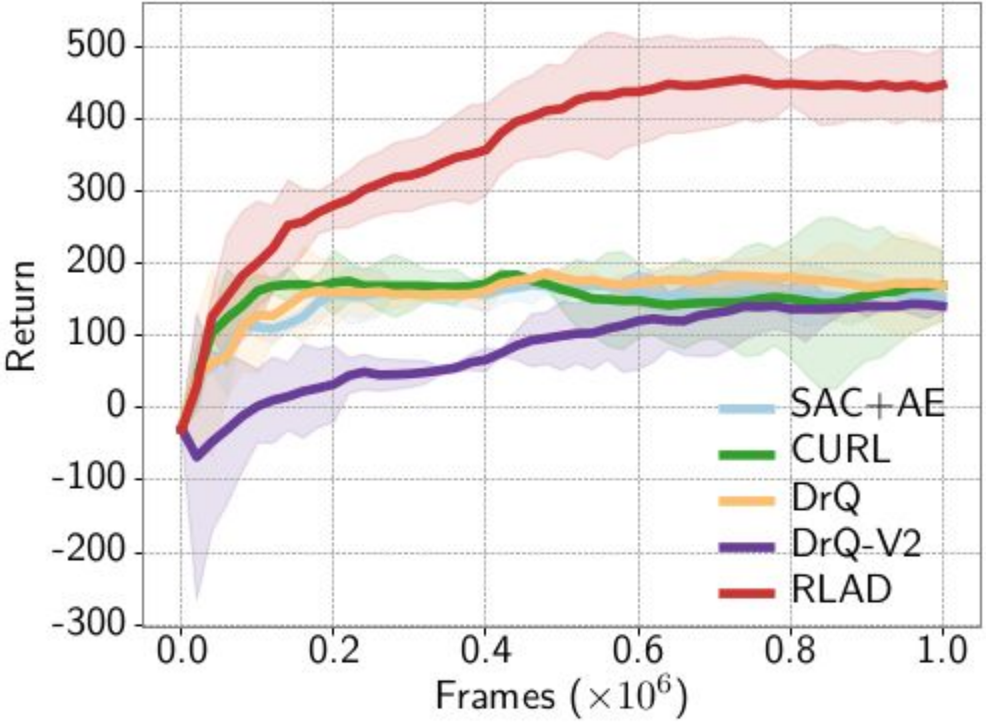
Regular

Dense

Setup of Experiments: **SOTA Methods**

- **SAC+AE**: AAAI 2021
- **CURL**: ICML 2020
- **DrQ**: ICLR 2021
- **DrQ-V2**: ICLR 2022

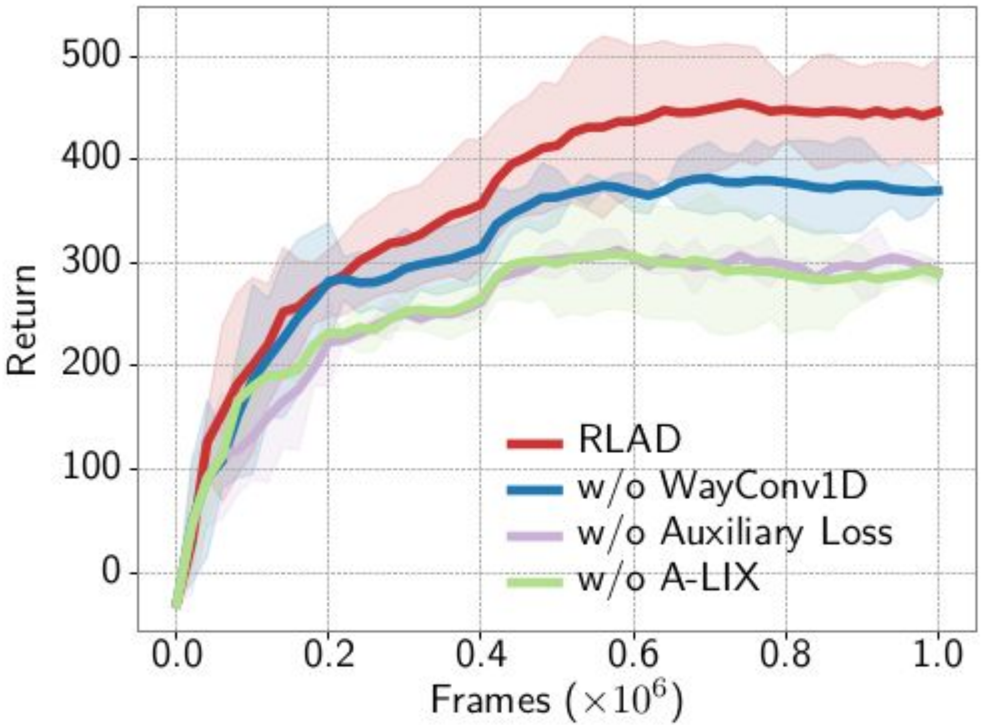
Comparison with SOTA: **Return**



Comparison with SOTA: **Success Rate (%)**

	Empty	Regular	Dense
SAC+AE	82	42	6
CURL	74	30	2
DrQ	**94**	42	10
DrQ-V2	10	8	0
RLAD	**94**	**62**	**32**

Ablation Study




4x
Training

Summary

- RLAD is the first algorithm that **learns simultaneously the encoder and the driving policy network using RL** in the domain of vision-based urban AD

- Although RLAD outperforms all RLfP methods in the urban AD domain, it is **not yet competitive** with state-of-the-art RL methods that decouple the training of encoder and the policy network [7] or that use expert demonstrations [8]

[7] M. Toromanoff et al., "End-to-end model-free reinforcement learning for urban driving using implicit affordances," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7151–7160, 2019.
[8] Chen et al., "Learning by Cheating". Proceedings of the Conference on Robot Learning, volume 100 of Proceedings of Machine Learning Research, 66–75. PMLR.

3.

RLfOLD: Reinforcement Learning from Online Demonstrations in Urban Autonomous Driving

Daniel Coelho, Miguel Oliveira, and Vitor Santos. "RLfOLD: Reinforcement Learning from Online Demonstrations in Urban Autonomous Driving." **Proceedings of the AAAI Conference on Artificial Intelligence**. Vol. 38. No. 10. **2024**, doi: 10.1609/aaai.v38i10.29049.

Motivation

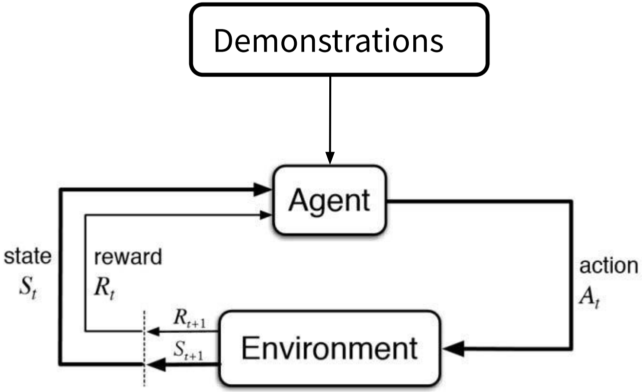
- How can we improve RLAD to outperform state-of-the-art methods on the NoCrash benchmark?

Motivation

- How can we improve RLAD to outperform state-of-the-art methods on the NoCrash benchmark?

By Integrating Expert Demonstrations

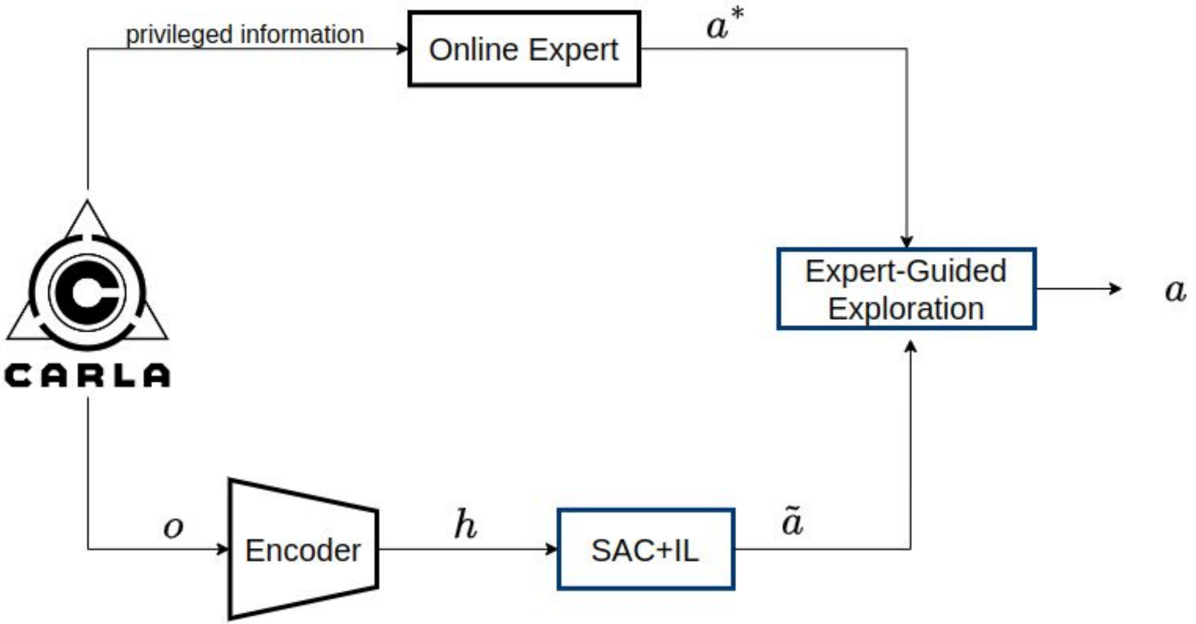
Reinforcement Learning from Demonstrations (RLfD)



✓
- High sample efficiency of IL
- Generalization of RL

✗
- Distribution gap between the demonstrations and the environment
- Complexity of integrating demonstrations within an RL framework

Learning Framework

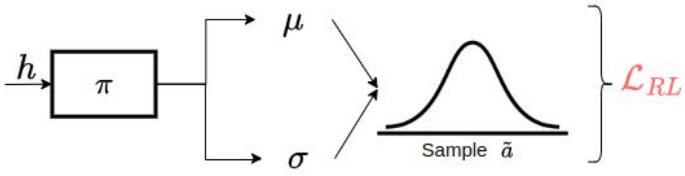


Soft Actor-Critic with Imitation Learning

- We propose a modified policy of the SAC algorithm that allows for the inclusion of the IL loss

Soft Actor-Critic with Imitation Learning

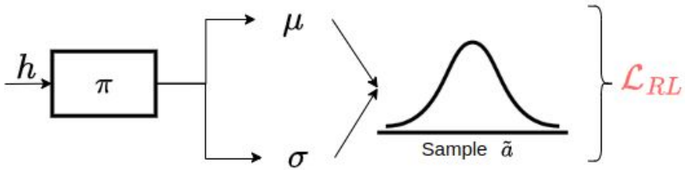
- We propose a modified policy of the SAC algorithm that allows for the inclusion of the IL loss



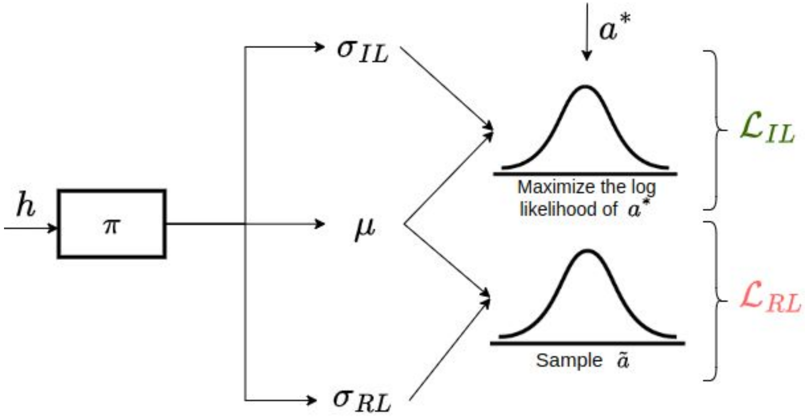
Traditional policy

Soft Actor-Critic with Imitation Learning

- We propose a modified policy of the SAC algorithm that allows for the inclusion of the IL loss



Traditional policy



Proposed policy

Soft Actor-Critic with Imitation Learning

- With different standard deviations, the algorithm can adapt to the **varying levels of uncertainty** in RL and IL

- It allows the RL component to explore the state-action space more broadly, while the IL component can focus on imitating the expert's behavior more closely

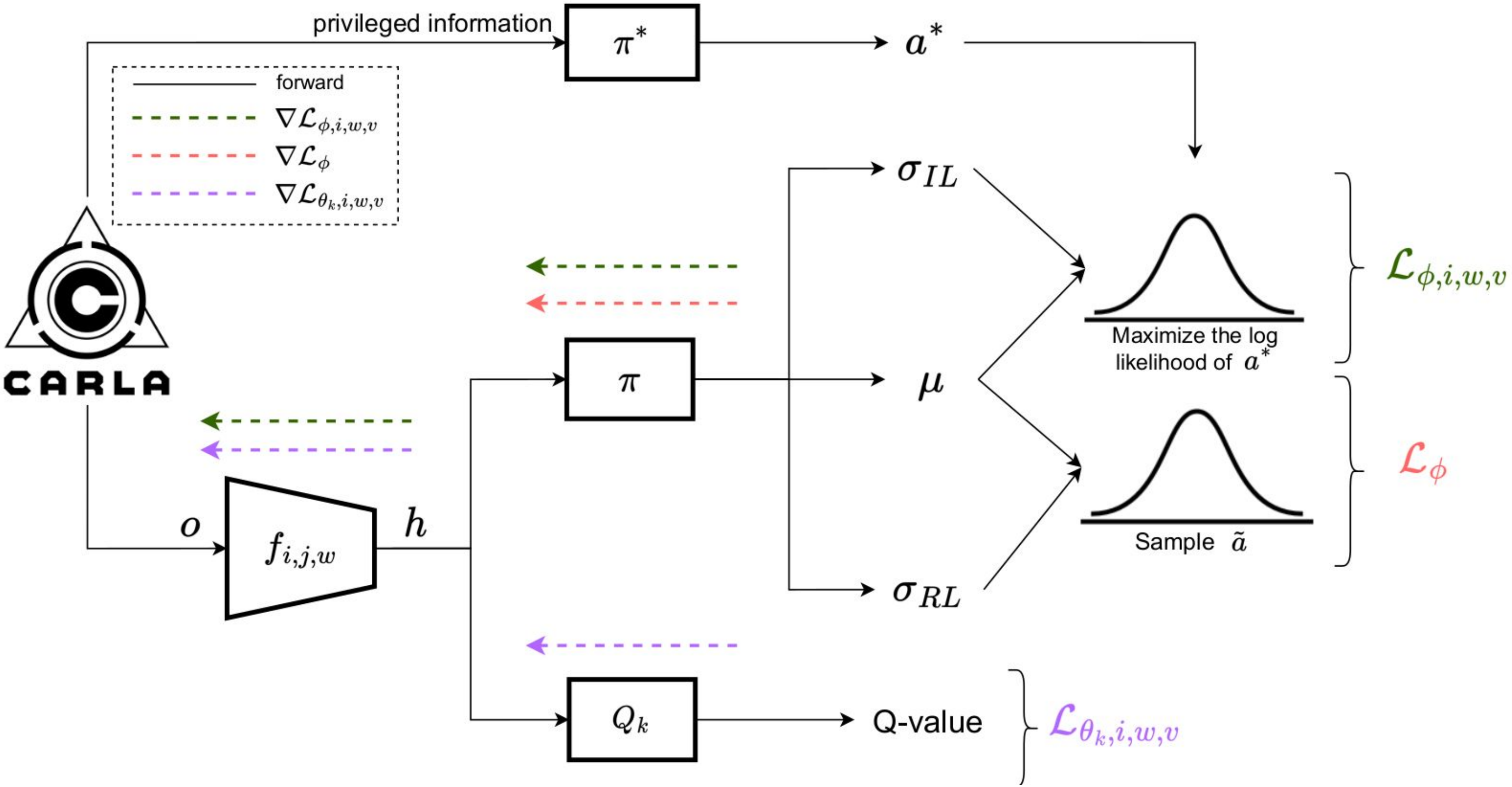
Expert-Guided Exploration Based on Uncertainty

- Rather than limiting the online expert's role to the IL loss, we also employ it to assist the exploration

- The idea is to use σ_{RL} as the uncertainty of the decision taken by the current policy

$$a = \begin{cases} \tilde{a} \text{ if } \sigma_{RL} < u \\ a^* \text{ otherwise} \end{cases}$$

RLfOLD


privileged information
π^*
a^*
forward
$\nabla\mathcal{L}_{\phi,i,w,v}$
$\nabla\mathcal{L}_{\phi}$
$\nabla\mathcal{L}_{\theta_k,i,w,v}$
CARLA
σ_{IL}
$\mathcal{L}_{\phi,i,w,v}$
Maximize the log likelihood of a^*
π
μ
$\mathcal{L}_{\phi}$
o
$f_{i,j,w}$
h
Sample $\tilde{a}$
σ_{RL}
Q_k
Q-value
$\mathcal{L}_{\theta_k,i,w,v}$

Algorithm 1: Reinforcement Learning from Online Demonstrations (RLfOLD)

Input: initial encoder parameters $f_{i,w,v}$, Q-function parameters Q_{θ_1}, Q_{θ_2}, policy parameters π_ϕ, entropy parameter α, empty replay buffer $\mathcal{D}$

```
Q_{θ̄_k} ← Q_{θ_k}, for k = 1,2
repeat
    Get observation o_t
    Compute expert action a*_t using π*
    Encode o_t into h_t using Equation 1
    Sample policy action ã_t ~ π_φ(· | h_t)
    Execute a_t according to Equation 6
    Get next observation o_{t+1} and reward r_t
    Store transition (o_t, a_t, a*_t, r_t, o_{t+1}) in D
    if o_{t+1} is terminal then
        Reset environment state
    end if
    if time to update then
        Randomly sample a batch of transitions, B = {(o_t, a_t, a*_t, r_t, o_{t+1})} from D
        Update Q_{θ1}, Q_{θ2} and f_{i,w,v} using Equation 2
        Update π_φ using Equation 4
        Update π_φ, and f_{i,w,v} using Equation 5
        Update α according to (Haarnoja et al. 2018)
        Update Q_{θ̄_k} with
            Q_{θ̄_k} ← (1 − ρ) Q_{θ̄_k} + ρ Q_{θ_k}, for k = 1,2
    end if
until convergence
```

$$\boldsymbol{h}_t = f_{i,w,v}\left(o_t\right) \tag{1}$$

$$a = \begin{cases} \tilde{a} \text{ if } \sigma_{RL} < u \\ a^* \text{ otherwise} \end{cases} \tag{6}$$

$$\mathcal{L}_{\theta_k,i,w,v} = \mathbb{E}_{\substack{o_t,a_t,o_{t+1}\sim\mathcal{D} \\ \tilde{a}_{t+1}\sim\pi_\phi(\cdot|\boldsymbol{h}_{t+1})}}\left[\left(Q_{\theta_k}\left(\boldsymbol{h}_t, a_t\right) - y\right)^2\right], \forall k \in \{1,2\} \tag{2}$$

$$\mathcal{L}_\phi = -\mathbb{E}_{\substack{o_t\sim\mathcal{D} \\ \tilde{a}_t\sim\pi_\phi(\cdot|\boldsymbol{h}_t)}}\left[\min_{k=1,2} Q_{\theta_k}\left(\boldsymbol{h}_t, \tilde{a}_t\right) - \alpha \log \pi\left(\tilde{a}_t \mid \boldsymbol{h}_t\right)\right] \tag{4}$$

$$\mathcal{L}_{\phi,i,w,v} = -\mathbb{E}_{o_t,a_t^*\sim\mathcal{D}}\left[\log p_\phi\left(a_t^* \mid \boldsymbol{h}_t\right)\right] \tag{5}$$

Setup of Experiments: **NoCrash Benchmark**

Town 01

Town 02

Setup of Experiments: **SOTA Methods**

- RL:
  - **IAs**: CVPR 2020
  - **CADRE**: AAAI 2022
- IL:
  - **CILRS**: ICCV 2019
  - **LBC**: CoRL 2020
- RLfD:
  - **GRIAD**: Robotics 2023
  - **WOR**: ICCV 2021

Comparison with SOTA: **Success Rate (%)**

			RL		IL		RLfD		
Task	Town	Weather	IAs	CADRE	CILRS	LBC	GRIAD*	WOR*	RLfOLD
Empty	train	train	85	95	97	89	98	98	**100**
Regular	train	train	85	92	83	87	98	**100**	94
Dense	train	train	63	82	42	75	94	**96**	90
Empty	test	train	77	92	66	86	94	94	**100**
Regular	test	train	66	78	49	79	**93**	89	92
Dense	test	train	33	61	23	53	78	74	**80**
Empty	train	test	-	94	**96**	60	83	90	**96**
Regular	train	test	-	86	77	60	87	**90**	84
Dense	train	test	-	76	39	54	83	**84**	74
Empty	test	test	-	78	66	36	69	78	**100**
Regular	test	test	-	72	56	36	63	82	**86**
Dense	test	test	-	52	24	12	52	**66**	**66**
Average	-	-	68	80	60	60	83	87	**89**

* Used 3 cameras as input.

Comparison with SOTA: **# of parameters and # of cameras**

	# of parameters	# of cameras
IAs	~30M	1
CADRE	~25M	1
CILRS	~22M	1
LBC	~22M	1
GRIAD	~14M	3
WOR	~22M	3
RLfOLD	~0.65M	1

Ablation Study

	Success rate %, ↑	Route completion %,↑	Collision pedestrian #/Km, ↓	Collision vehicle #/Km, ↓	Collision layout #/Km, ↓	Agent blocked #/Km, ↓
RL baseline	52±4	98±3	1.03±0.34	1.40±0.11	0.26±0.05	0.36±0.13
RLfOLD w/o two SDs	64±10	90±6	0.33±0.13	0.53±0.09	0.15±0.09	4.45±1.43
RLfOLD w/o uncertainty (p=0.0)	72±2	96±3	0.14±0.04	0.48±0.03	0.12±0.03	3.99±0.47
RLfOLD w/o uncertainty (p=0.3)	80±3	91±1	0.30±0.04	0.45±0.06	**0.00**±0.00	2.76±0.91
RLfOLD	**86**±4	**99**±2	**0.09**±0.03	**0.32**±0.04	0.09±0.03	**0.15**±0.08

Summary

- RLfOLD introduces a seamless integration of IL and RL by **leveraging online demonstrations**, a **dual standard deviation policy network**, and an **uncertainty-based technique** guided by an online expert to enhance the exploration process

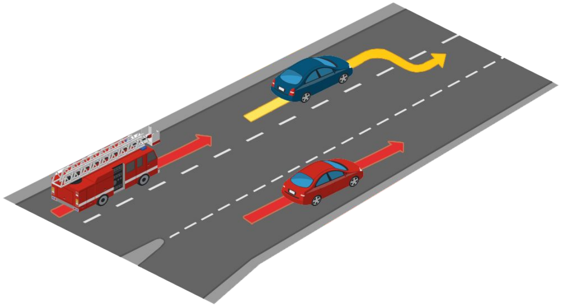
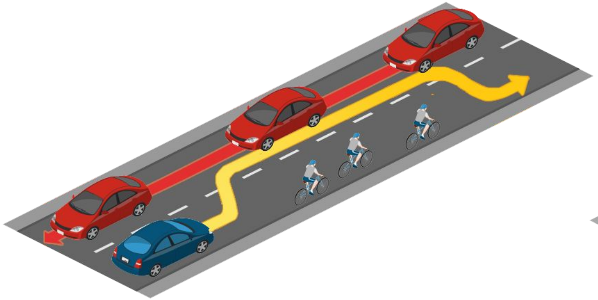
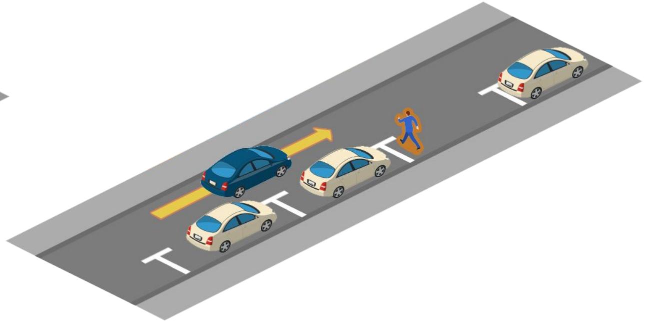
- Even with a significantly smaller encoder and a single-camera setup, RLfOLD surpasses all state-of-the-art methods on the NoCrash benchmark

4.

PRIBOOT: A New Data-Driven Expert for Improved Driving Simulations

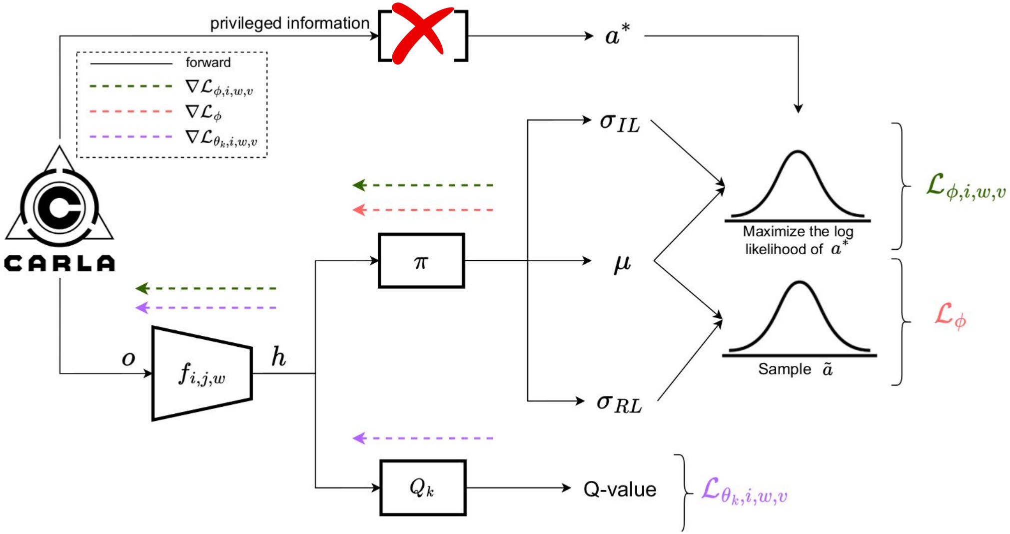
Motivation

- After achieving top performance on the NoCrash benchmark, we advanced to the more recent and challenging CARLA benchmark: **Leaderboard 2.0**

Motivation

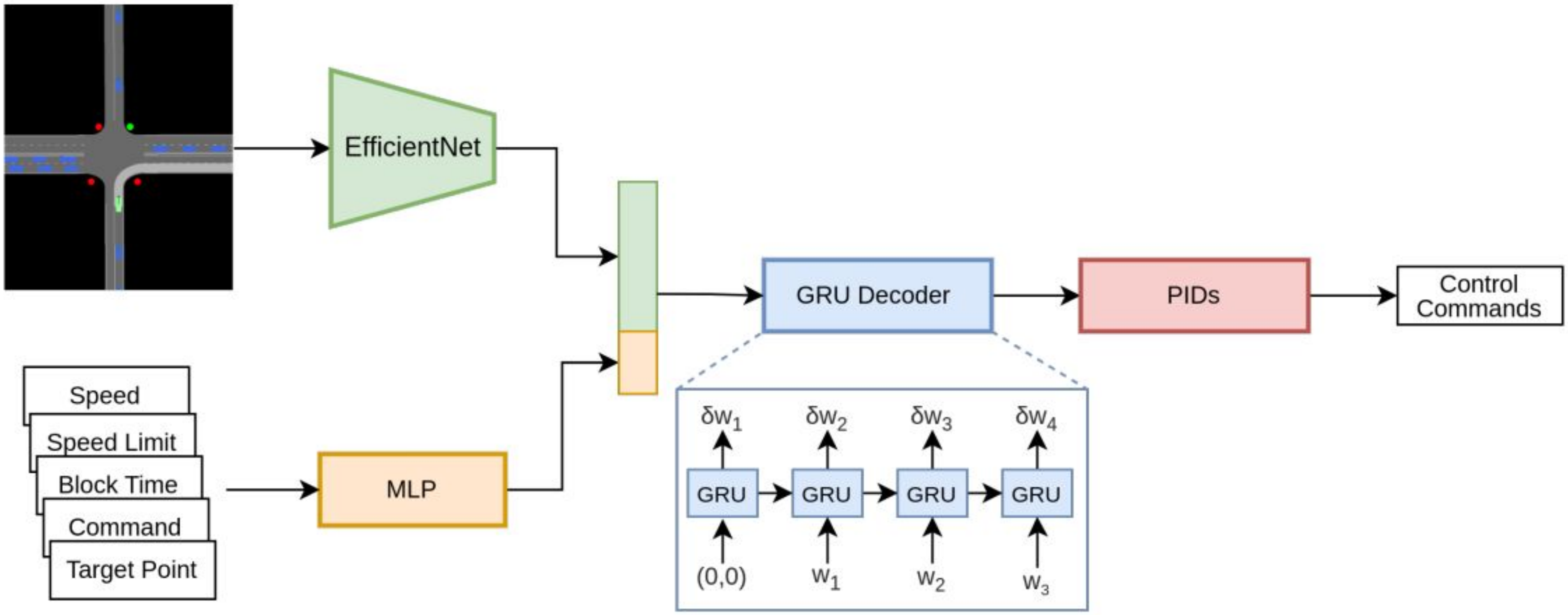
- The objective was to use **RLfOLD** in **Leaderboard 2.0**; however, there was no online expert working effectively



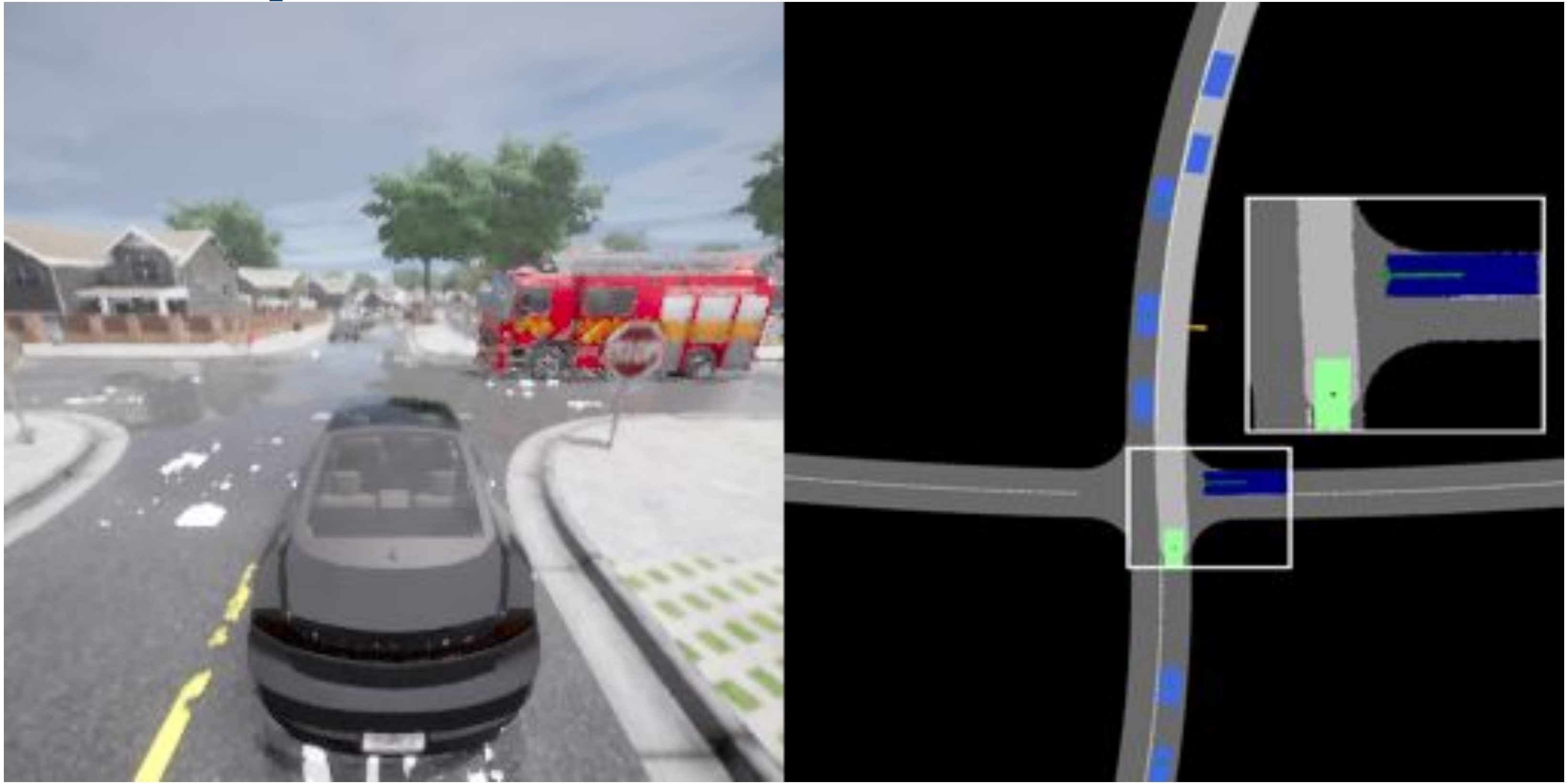
PRIBOOT (**Pri**vileged Information **Boot**strapping)

- This work proposes PRIBOOT, the first functional online expert for the Leaderboard 2.0
- CARLA provides **human driving logs**, which, while insufficient for models requiring sensor inputs, become valuable when combined with **privileged information**
- PRIBOOT is capable of navigating the demanding scenarios presented in Leaderboard 2.0, subsequently enabling the generation of extensive datasets or providing online demonstrations

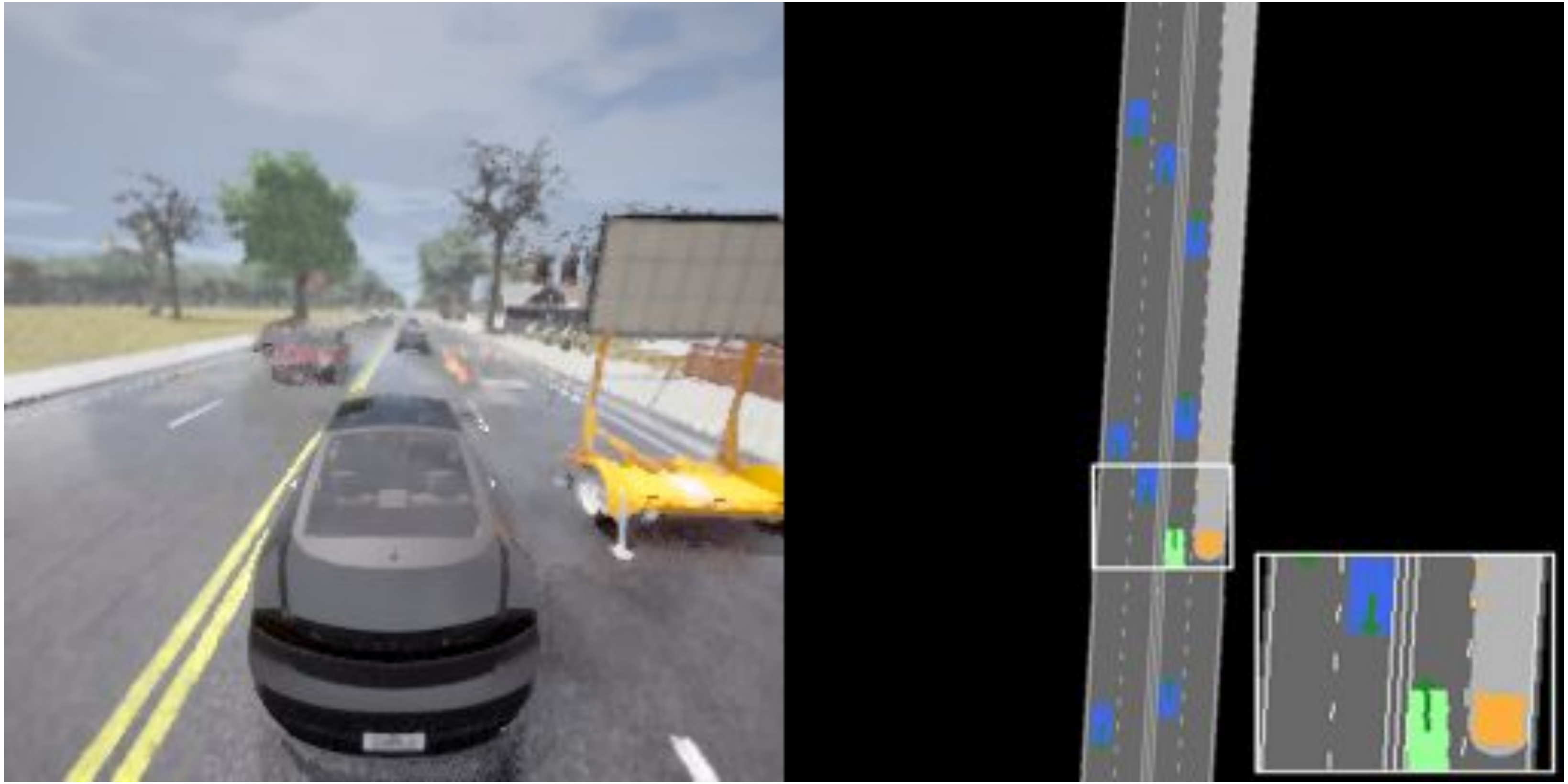
Architecture



BEV: Example 1



BEV: Example 2



Driving Score (DS)

- DS is the main metric to evaluate the performance of models in Leaderboard 2.0
- However, this metric biases against longer routes due to its cumulative penalty for infractions

$$\mathbf{DS} = \mathbf{RC} \cdot \prod_{i=1}^{q} p_i^{n_i}$$

Driving Score (DS)

- For instance, let's consider an agent with an average infraction rate of 0.2 collisions per km (penalty=0.6)

- Considering that the route completion is 100% we have very different results if we test this agent in a 5 km route or 10 km route
  - 5 km: DS= 1 * 0.6^1 = 0.6
  - 10km: DS = 1* 0.6*2 = 0.36

$$\mathbf{DS} = \mathbf{RC} \cdot \prod_{i=1}^{q} p_i^{n_i}$$

Infraction Rate Score (IRS)

- To promove fairness, we introduce IRS. This metric accounts for the infraction rate per kilometer, adjusting for route length and providing a balanced evaluation

$$\mathbf{IRS} = \mathbf{RC} \cdot \prod_{i=1}^{q} e^{-\lambda \cdot \frac{n_i}{L} \cdot (1 - p_i)}$$

Comparison with Baseline

		DS ↑ %	IRS ↑ %	RC ↑ %	IP ↑ %	C.P ↓ #/Km	C.V ↓ #/Km	C.L ↓ #/Km	R.L ↓ #/Km	Stop ↓ #/Km	O.R ↓ #/Km	R.D ↓ #/Km	Block ↓ #/Km	Y.E ↓ #/Km	S.T ↓ #/Km	M.S ↓ #/Km
Town12	Autopilot	1.22	0.51	5.97	0.26	1.26	4.59	0.58	0.11	1.84	0.62	0.66	1.26	**0.00**	0.34	**0.00**
	PRIBOOT	**22.80**	**42.75**	**76.46**	**0.30**	**0.00**	**0.31**	**0.06**	**0.01**	**0.02**	**0.05**	**0.00**	**0.06**	0.04	**0.03**	0.11
Town13	Autopilot	0.99	0.22	5.55	0.20	0.83	3.06	0.83	**0.00**	0.02	0.35	0.69	0.69	**0.00**	0.10	**0.00**
	PRIBOOT	**18.84**	**46.97**	**74.29**	**0.24**	**0.01**	**0.34**	**0.05**	**0.00**	**0.01**	**0.05**	**0.00**	**0.04**	0.02	**0.02**	0.06

Abbreviation	Full Name
DS	Driving Score
IRS	Infraction Rate Score
RC	Route Completion
IP	Infraction Penalty
C.P	Collisions Pedestrians
C.V	Collisions Vehicles
C.L	Collisions Layout
R.L	Red Light Infractions
Stop	Stop Sign Infractions
O.R	Off-road Infractions
R.D	Route Deviation
Block	Agent Blocked
Y.E	Yield Emergency Infractions
S.T	Scenario Timeouts
M.S	Min Speed Infractions

Demonstration of PRIBOOT' Driving

Demonstration of PRIBOOT' Driving

Summary

- PRIBOOT is a system that utilizes **privileged information** alongside **limited human driving logs** to establish the first expert in the CARLA Leaderboard 2.0

- PRIBOOT enables researchers to **generate extensive datasets**, potentially resolving data availability issues in this benchmark.

5.

Conclusions

Conclusions

- This thesis presents a significant progress in end-to-end AD for urban environments, with a focus on RL
- Introduced **RLAD**, **RLfOLD**, and **PRIBOOT** leveraging RL and IL to achieve state-of-the-art results in the NoCrash benchmark, and to introduce the first online expert of Leaderboard 2.0

Contributions

- **A Review of End-to-End Autonomous Driving in Urban Environments**
  - IEEE Access, 2022

- **RLAD: Reinforcement Learning From Pixels for Autonomous Driving in Urban Environments**
  - IEEE Transactions on Automation Science and Engineering, 2023

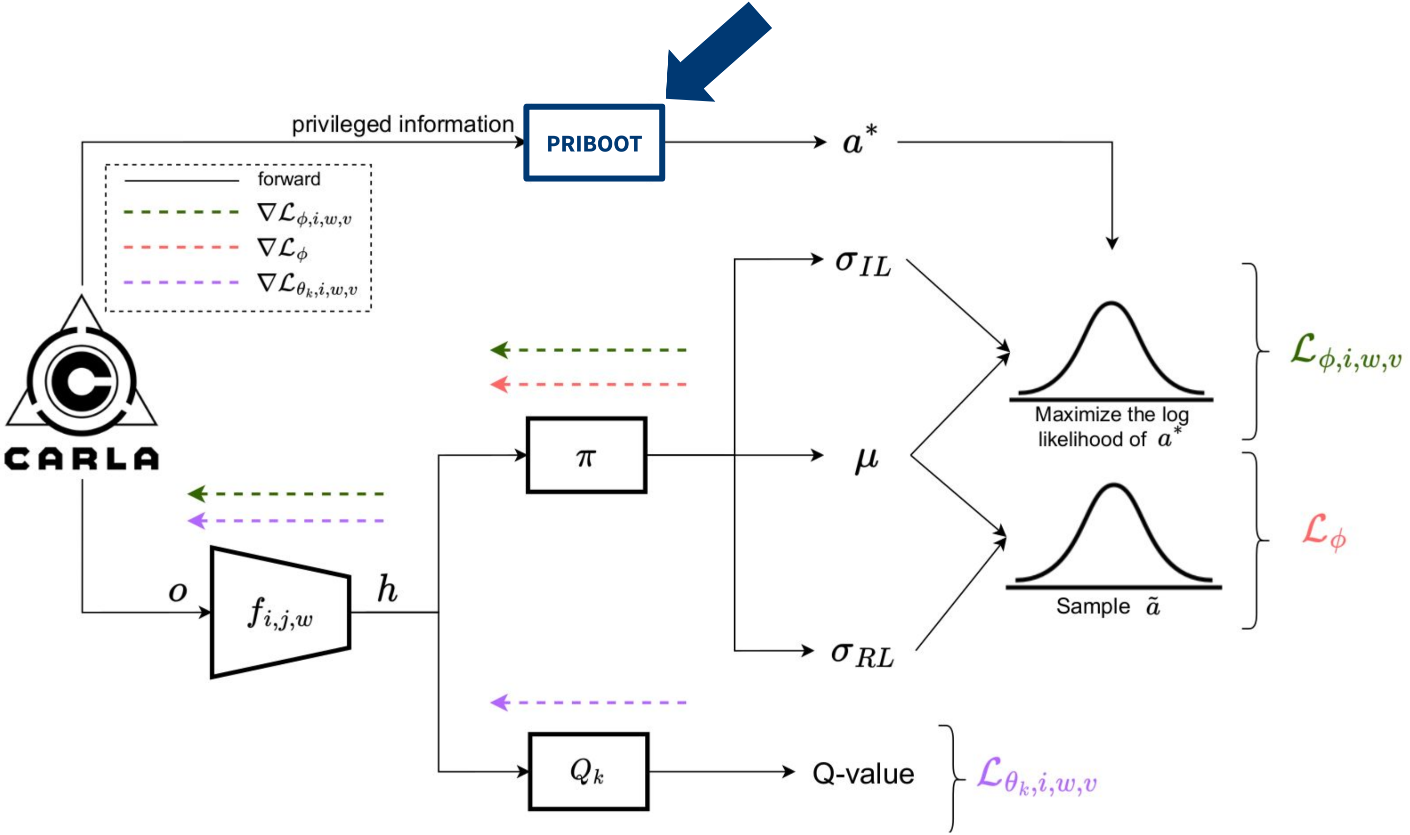
- **RLfOLD: Reinforcement Learning from Online Demonstrations in Urban Autonomous Driving**
  - Proceedings of the AAAI Conference on Artificial Intelligence, 2024

- **PRIBOOT: A New Data-Driven Expert for Improved Driving Simulations**
  - Submitted at IEEE Robotics and Automation Letters

Research Objectives

- Development of End-to-End RL Architectures for AD Systems in Urban Environments → **RLAD**
- Integration of Expert Demonstrations in an End-to-End RL Architecture for AD Systems → **RLfOLD**
- Development of a Data-Driven Expert Agent for Improved Driving Simulations → **PRIBOOT**

Future Work: **RLfOLD + PRIBOOT**



universidade de aveiro

ieeta

fct Fundação para a Ciência e a Tecnologia

CVC Computer Vision Center

End-to-End Reinforcement Learning for Autonomous Driving in Urban Environments

Thank you for your time!