CIRL: Controllable Imitative Reinforcement Learning for Vision-based Self-driving

ECCV2018 勉強会
B4 水谷純暉

内容

- 概要
- 関連手法
- 提案手法
- 比較
- 汎化能力
- 実利用に向けて
- まとめ

概要

- 模倣学習と強化学習を組み合わせて、高精度な自動運転モデルを学習
- 運転シミュレータ(CARLA)上で学習、評価
- リアルシーンにも適用

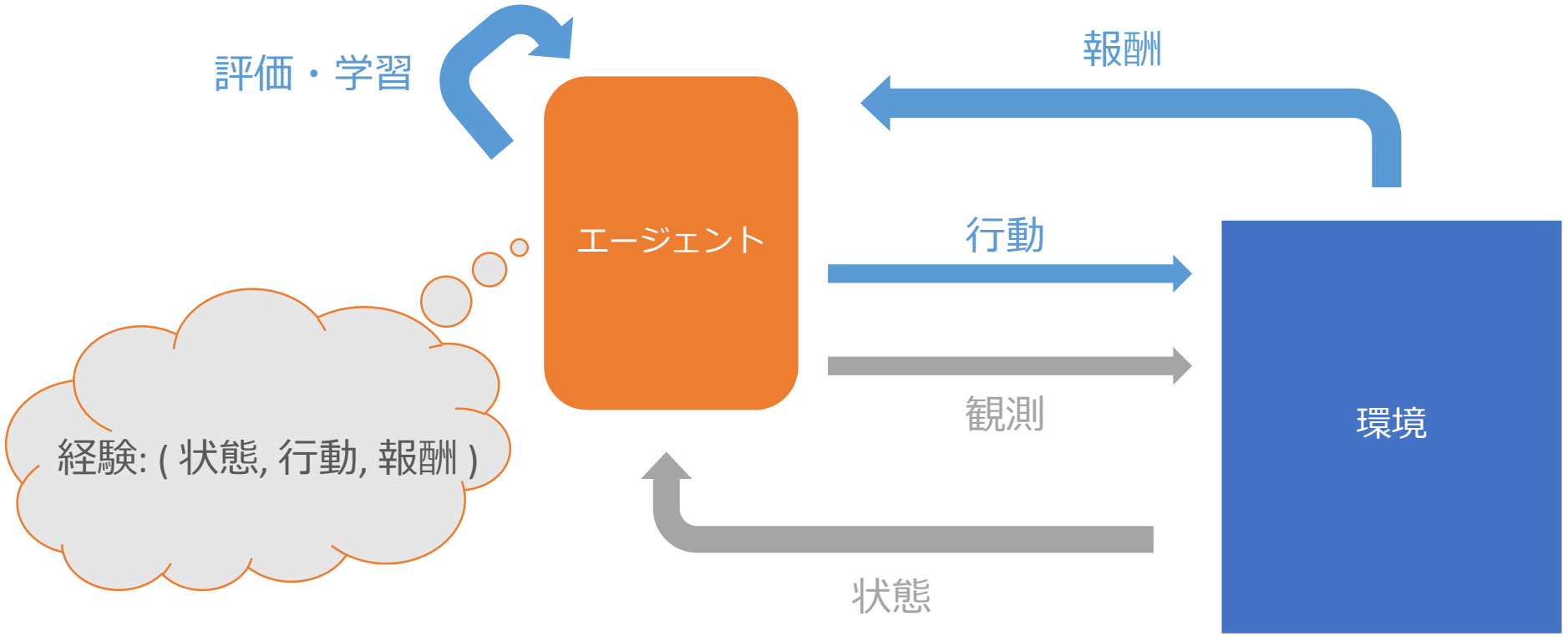
関連手法

関連手法

- 強化学習 (RL)
  - Actor– Critic
  - A3C
- 模倣学習 (IL)
- モジュール式パイプライン (MP)
  - 様々な手法の組み合わせ

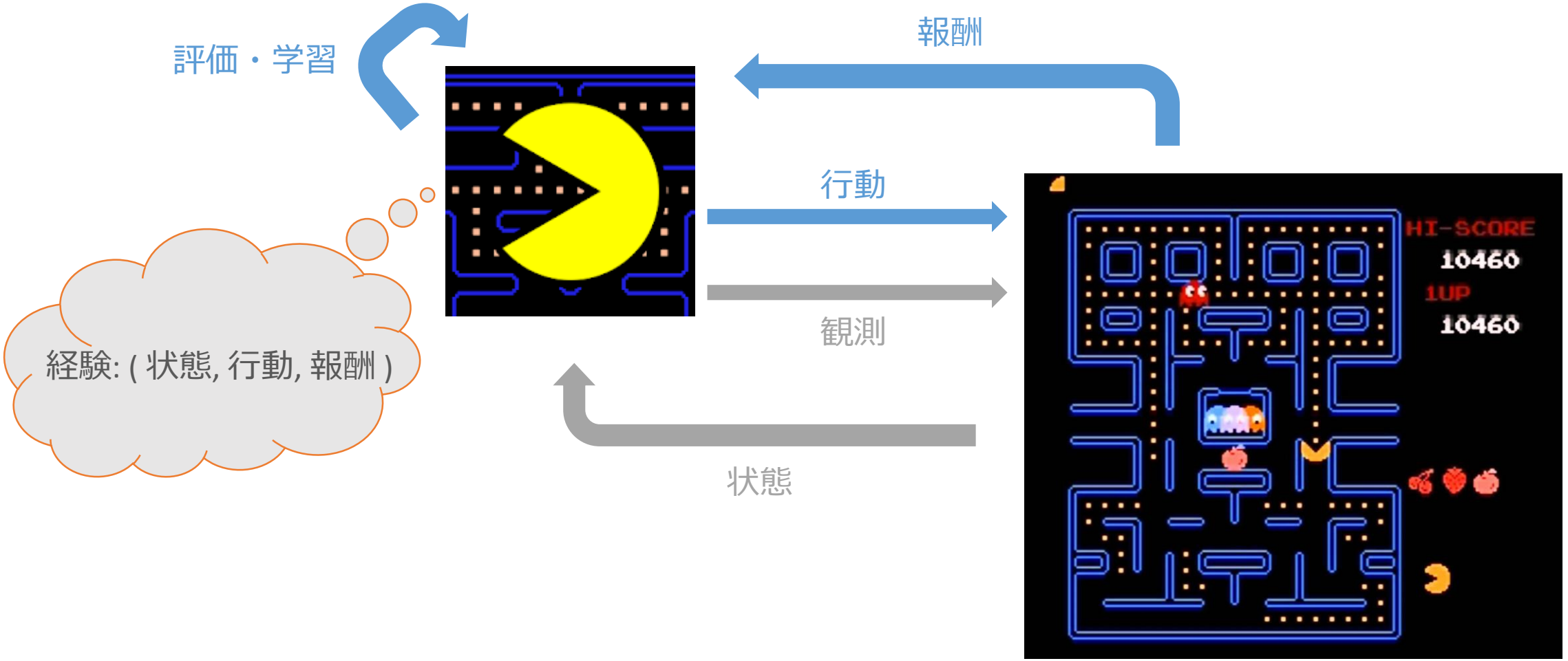
関連手法　強化学習

■ 強化学習とは
- 報酬に基づいた行動を取り、最適な行動則を試行錯誤して獲得していく学習方法
- ある状態でどんな行動をすれば、どのような報酬がもらえるか経験



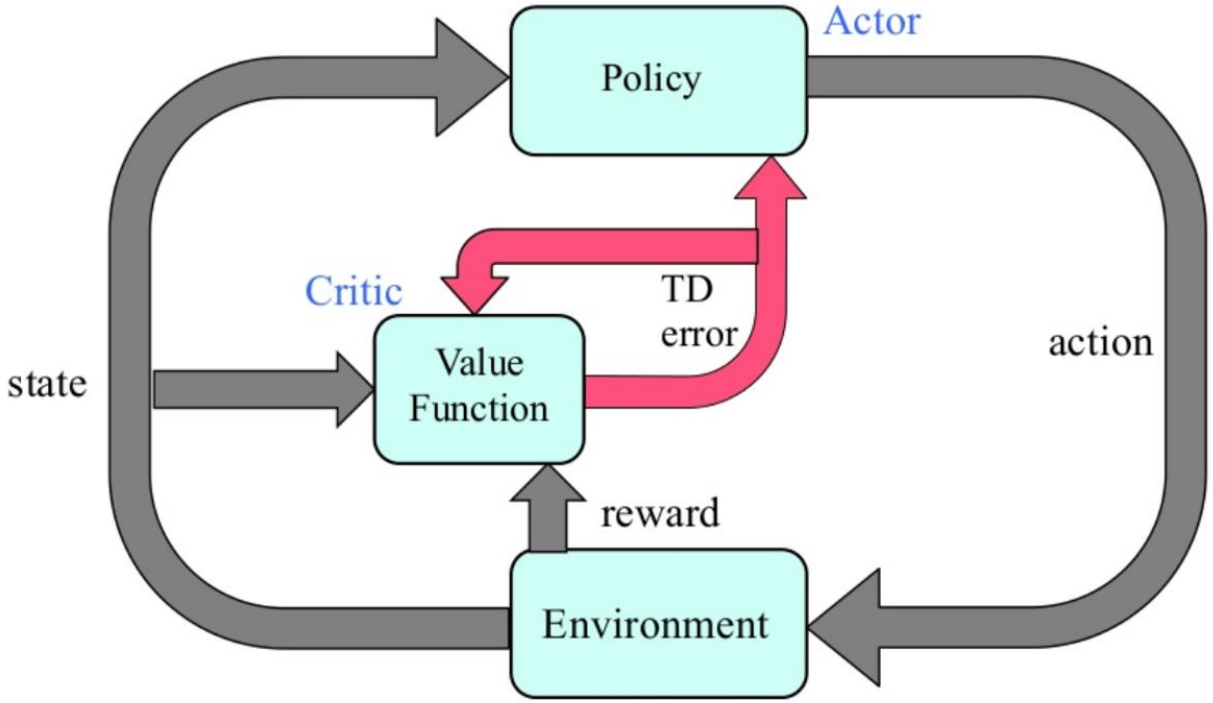
関連手法　強化学習

■ 強化学習とは
- 報酬に基づいた行動を取り、最適な行動則を試行錯誤して獲得していく学習方法
- ある状態でどんな行動をすれば、どれくらい報酬がもらえるか経験



関連手法　強化学習

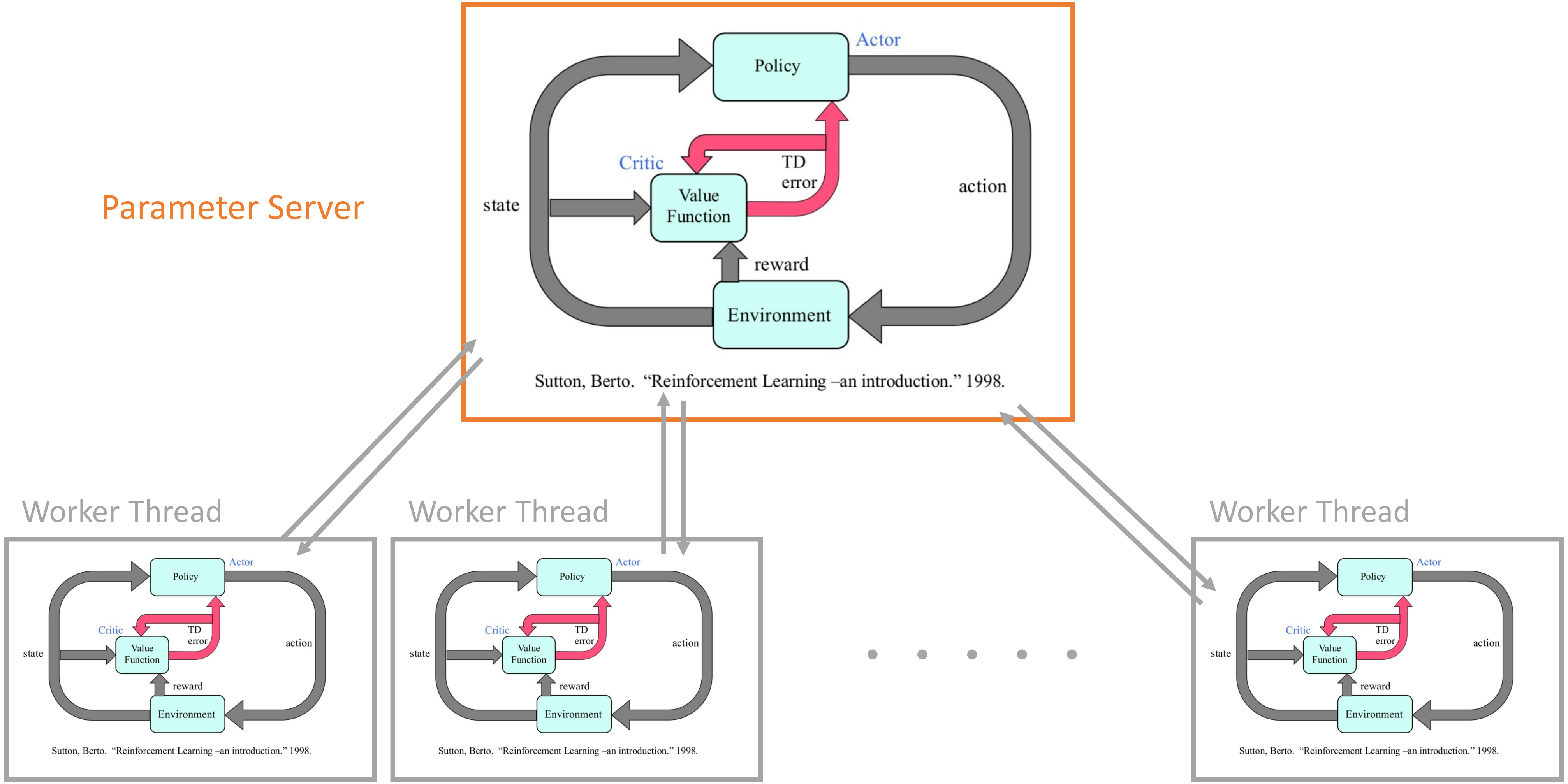
- Actor-Critic
  - 行動選択と評価のネットワークを分離
  - Actor (行動者)
    - エージェントとして行動を実行するネットワーク
  - Critic (評価者)
    - 報酬に基づいてActorの行動を評価するネットワーク



Sutton, Berto. “Reinforcement Learning –an introduction.” 1998.

関連手法　強化学習

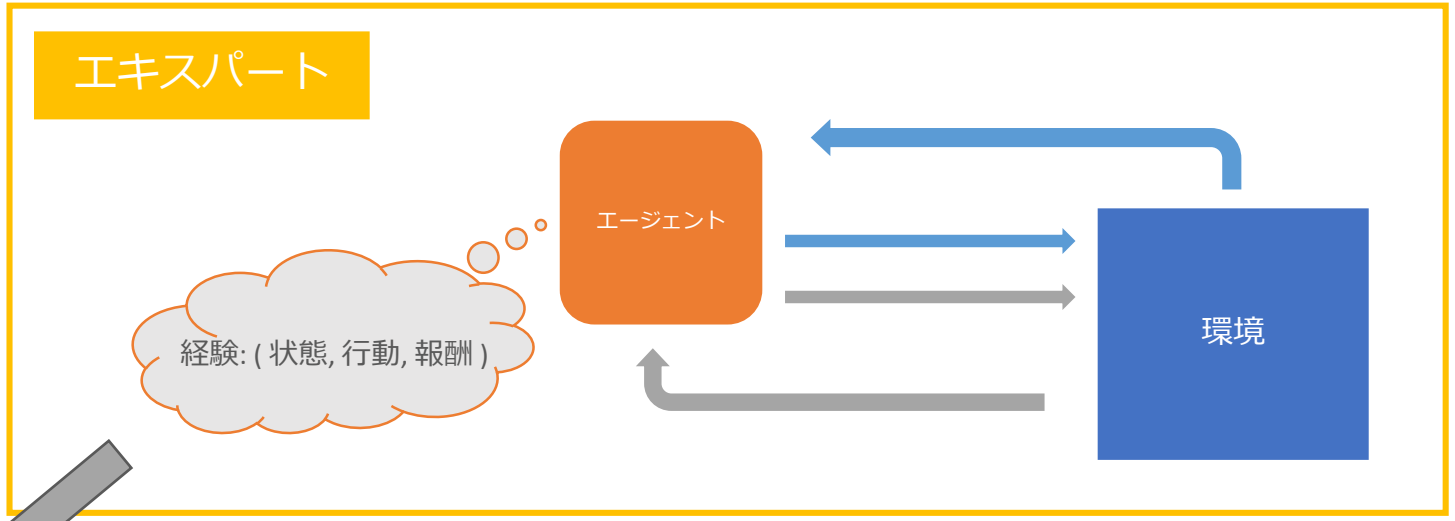
- A3C
  - Asynchronous Advantage Actor-Critic



関連手法　模倣学習

- 強化学習との違い
  - 強化学習
    - 自ら行動して得た経験で学習
  - 模倣学習
    - エキスパートの経験で学習
    - 教師あり学習

学習データ

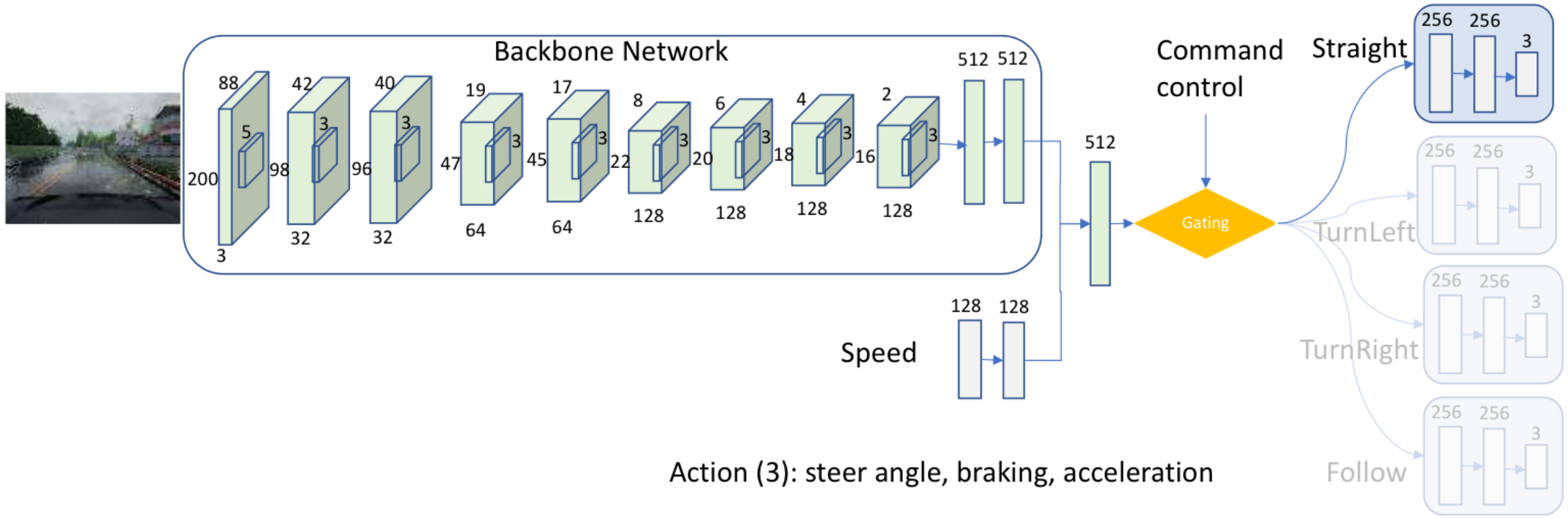


教師あり学習

π_θ(行動 | 状態)

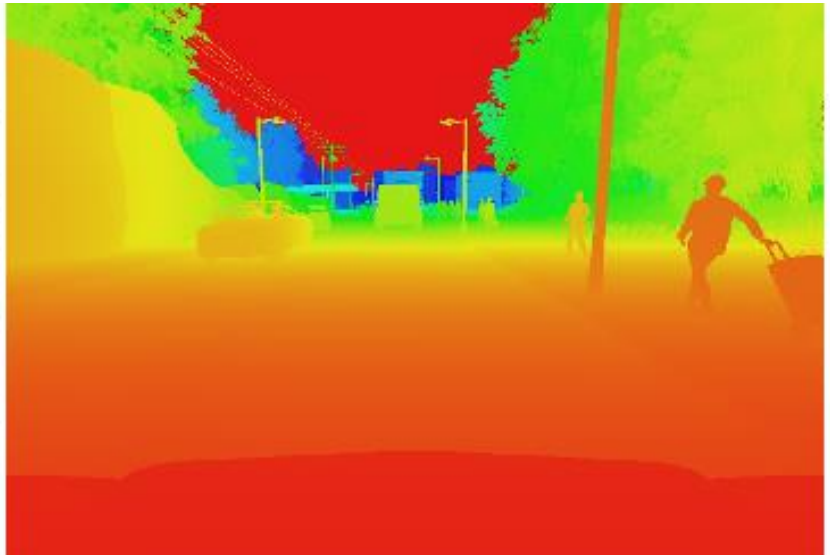
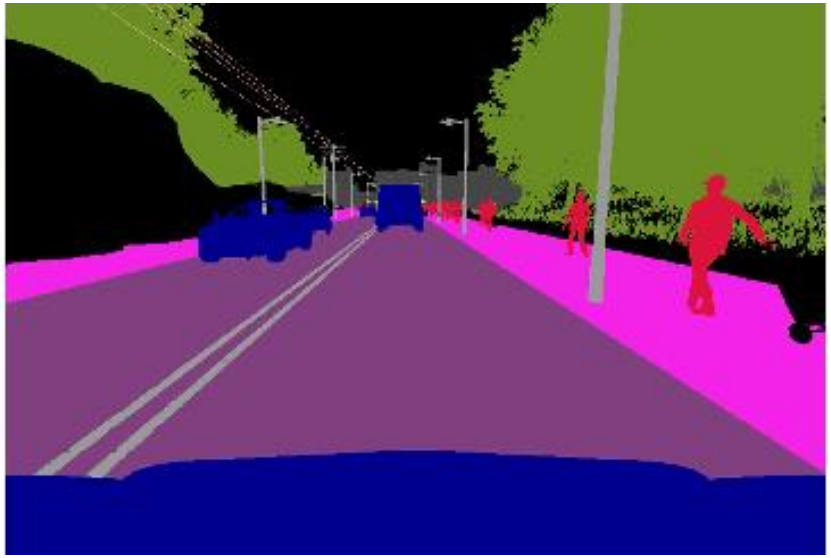
関連手法　模倣学習

- Conditional Imitation Learning
  - Command Control (Gating)
    - Follow (道なりに進む)
    - Straight (交差点を直進)
    - Left (交差点を左折)
    - Right (交差点を右折)



関連手法　モジュール式パイプライン

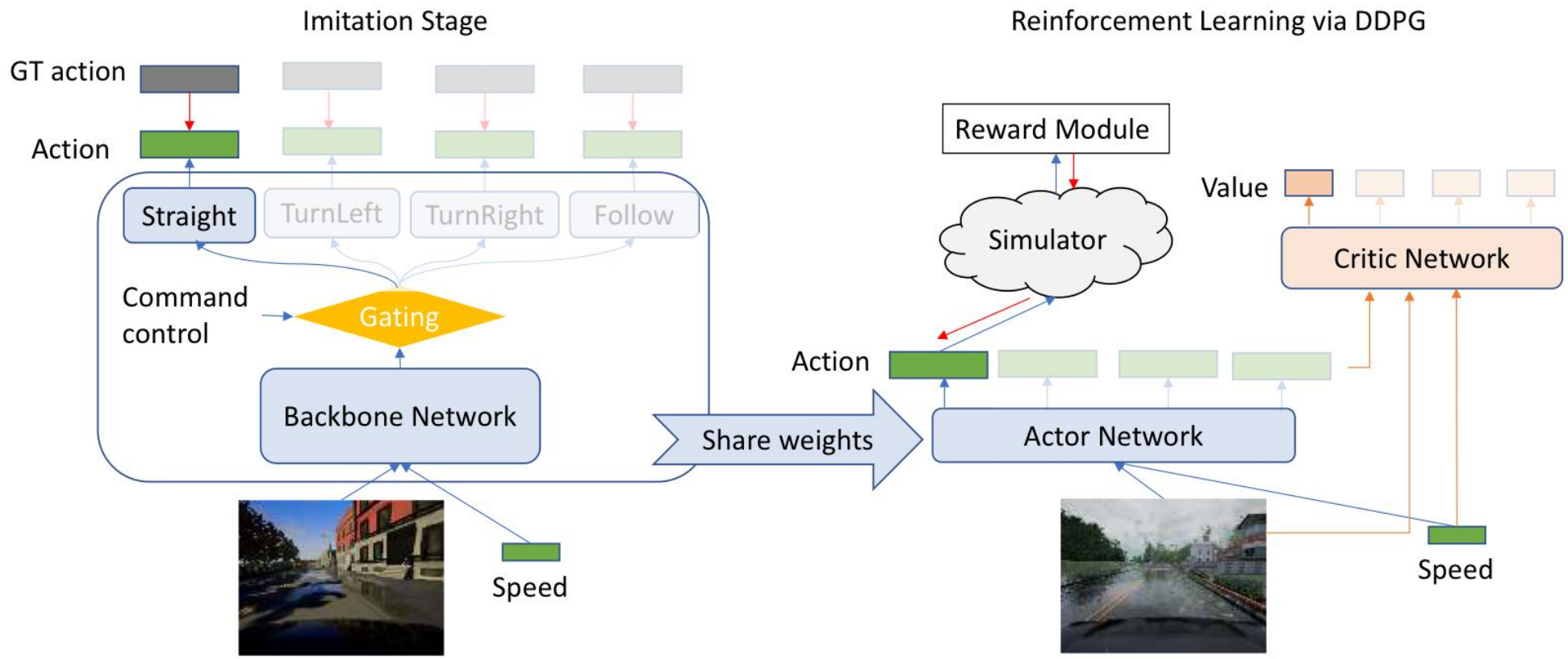
- Semantic Segmentation
  - 道路、歩道、車線、静的物体、動的物体　に分割
  - 分割情報に基づいた手作業の規則で運転

提案手法

提案手法

- CIRL: Controllable Imitative Reinforcement Learning
  - 模倣学習した後に強化学習
- メリット
  - サンプルの複雑さを大幅に低減
  - 学習時間の大幅短縮

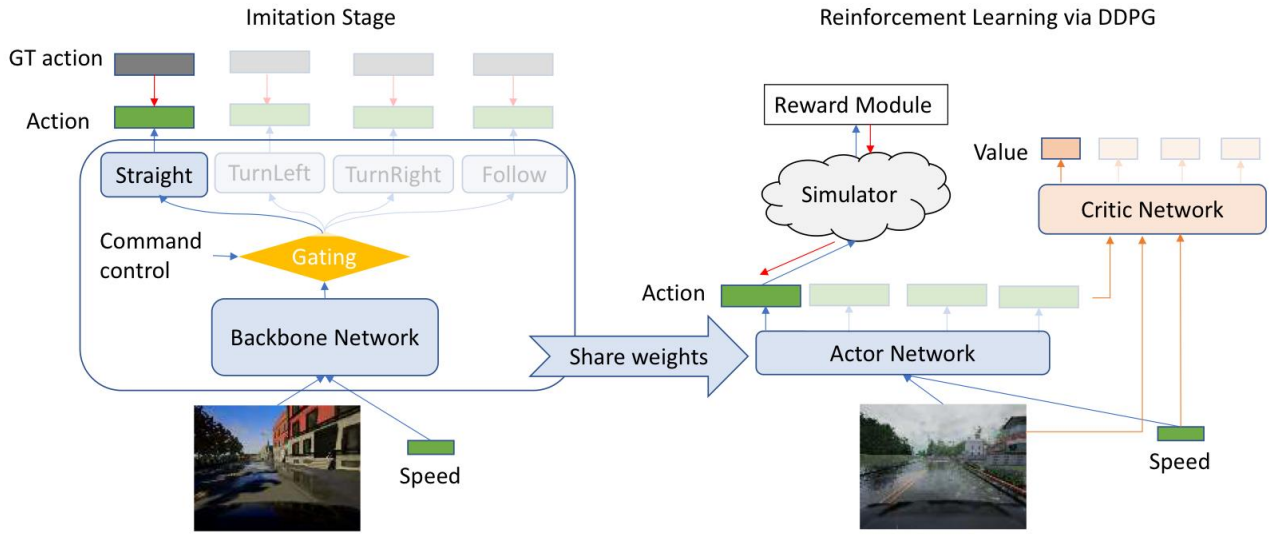


提案手法　データセット

- 使用データ
  - 概要
    - 町１、町２
    - 新天候セット１、新天候セット２
  - 学習
    - 町１
    - 新天候セット１
      - 晴れ、晴れの日の出、昼間の雨、雨の後の昼間
  - テスト
    - 町２
    - 新天気セット２
      - 正午の曇り、正午の雨、曇りの日の出、日の出時の激しい雨

提案手法　行動

- Gating
  - Follow (道なりに進む)
  - Straight (交差点を直進)
  - Left (交差点を左折)
  - Right (交差点を右折)
- Action
  - Steering angle
  - Acceleration
  - Braking



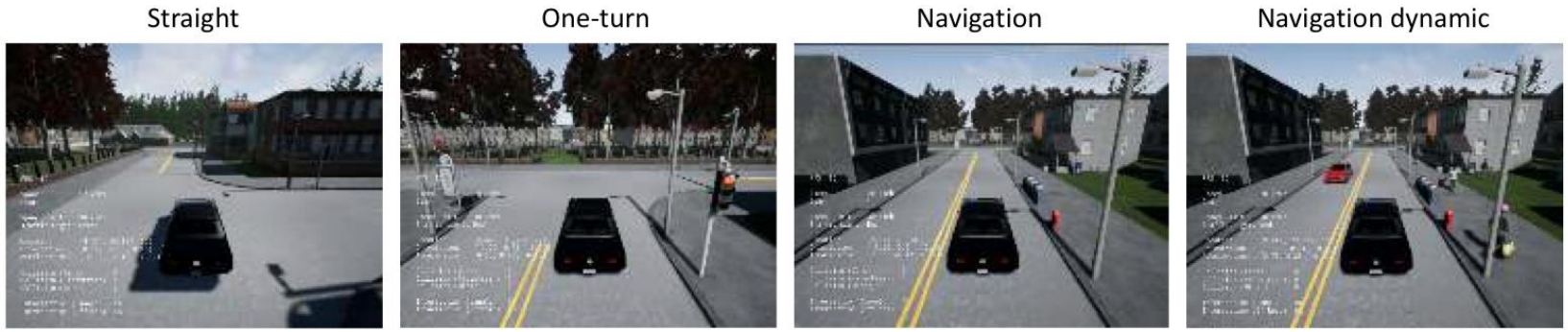


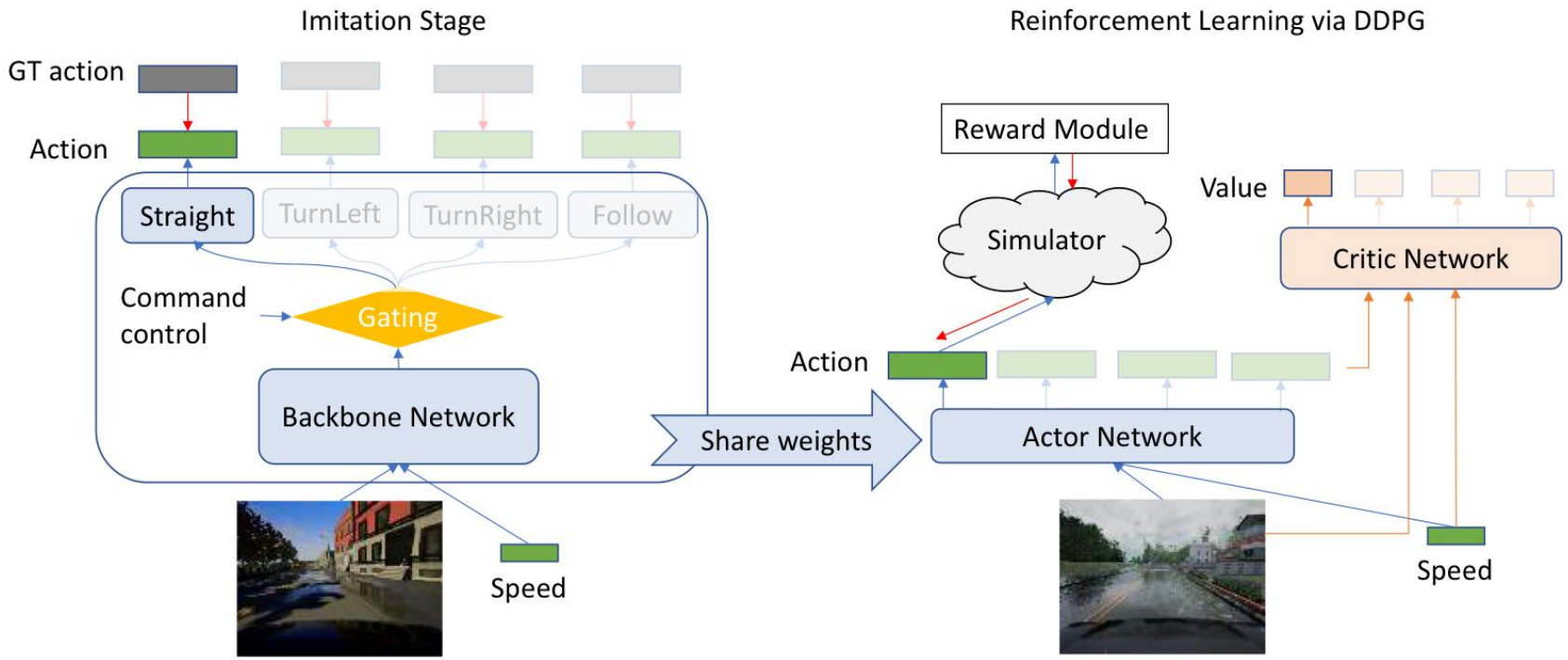
Fig. 5. Illustrated observations of four different tasks in the bird view.

提案手法　学習

- 模倣学習
  - 人間が操作した動画を学習データとして使用

$$\min_{\theta^I} \sum_{i}^{N} \sum_{t}^{T_i} \mathcal{L}(F(I_{i,t}, G(c_{i,t}), s_{i,t}), \mathbf{a}_{i,t})$$

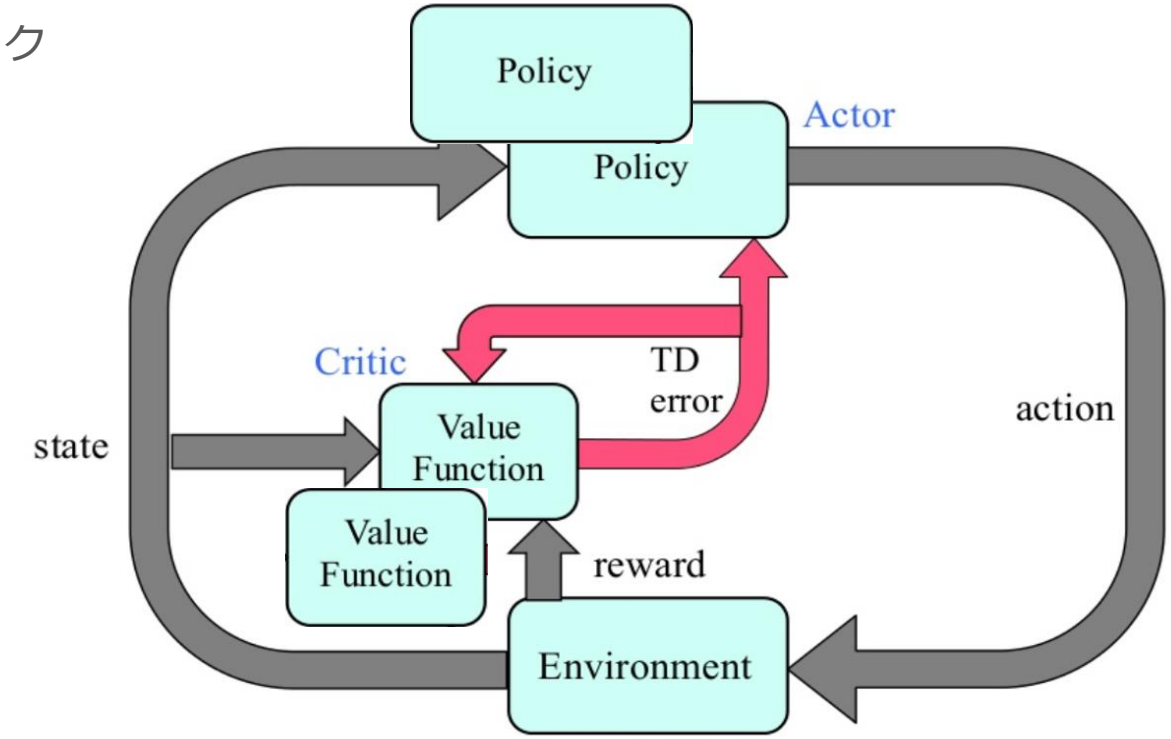
$$\mathcal{L}(\hat{\mathbf{a}}_{i,t}, \mathbf{a}_{\mathbf{i},\mathbf{t}}) = ||\hat{a}^s_{i,t} - a^s_{i,t}||^2 + ||\hat{a}^a_{i,t} - a^a_{i,t}||^2 + ||\hat{a}^b_{i,t} - a^b_{i,t}||^2$$



提案手法　Actor - Critic

- Actor-Critic
  - 行動選択と評価のネットワークを分離
  - Actor (行動者)
    - エージェントとして行動を実行するネットワーク
  - Critic (評価者)
    - 報酬に基づいてActorの行動を評価するネットワーク

- Target Network
  - 最適化目標を一定期間固定
  - Actor, Critic ネットワークそれぞれが安定した学習を行うために必要



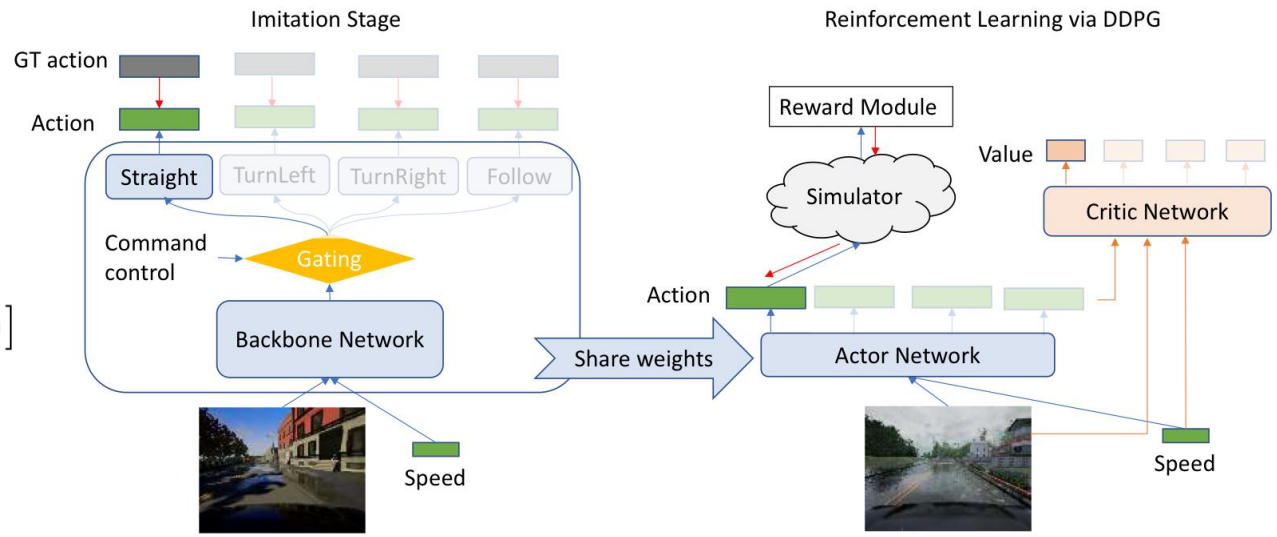
Sutton, Berto. “Reinforcement Learning –an introduction.” 1998.

強化学習

— DDPG: Deep Deterministic Policy Gradient

$$\mathcal{L}(\theta^Q) = \mathbf{E}_{(o,\mathbf{a},r,o')\sim D}[R - Q(o,\mathbf{a}|\theta^Q)]^2$$

$$\nabla_{\bar{\theta^\pi}} J(\bar{\theta^\pi}) \approx \mathbf{E}_{o,\mathbf{a}\sim D}[\nabla_{\mathbf{a}} Q(o,\mathbf{a}|\theta^Q)_{|\mathbf{a}=\pi(o,\theta^Q)} \nabla_{\theta_\pi} \pi(o|\bar{\theta^\pi})]$$



$$r_v(c) = \begin{cases} \min(25, v) & \text{if c for Follow} \\ \min(35, v) & \text{if c for Straight} \\ v & \text{if } v \leq 20\text{, c for TurnLeft and TurnRight} \\ 40 - v & \text{if } v > 20\text{, c for TurnLeft and TurnRight} \end{cases}$$

$$r_s(c) = \begin{cases} -15 & \text{if } s \text{ is in opposite direction with c for TurnLeft and TurnRight} \\ -20 & \text{if } |s| > 0.2\text{, c for Straight.} \end{cases}$$

$$r = R(o,\mathbf{a}) = r_s(c) + r_v(c) + r_r + r_o + r_d$$

比較

比較

■ タスク別の成功率

Straight

One-turn

Navigation

Navigation dynamic

Task	Training conditions				New town				New weather				New town/weather			
	MP	IL	RL	CIRL	MP	IL	RL	CIRL	MP	IL	RL	CIRL	MP	IL	RL	CIRL
Straight	98	95	89	**98**	92	97	74	**100**	100	98	86	**100**	50	80	68	**98**
One turn	82	89	34	**97**	61	59	12	**71**	**95**	90	16	94	50	48	20	**82**
Navigation	80	86	14	**93**	24	40	3	**53**	**94**	84	2	86	47	44	6	**68**
Nav. dynamic	77	**83**	7	82	24	38	2	**41**	**89**	82	2	80	44	42	4	**62**

比較　RL

- 結果
  - ほぼ全ての比較手法、タスクにおいて低い成功率

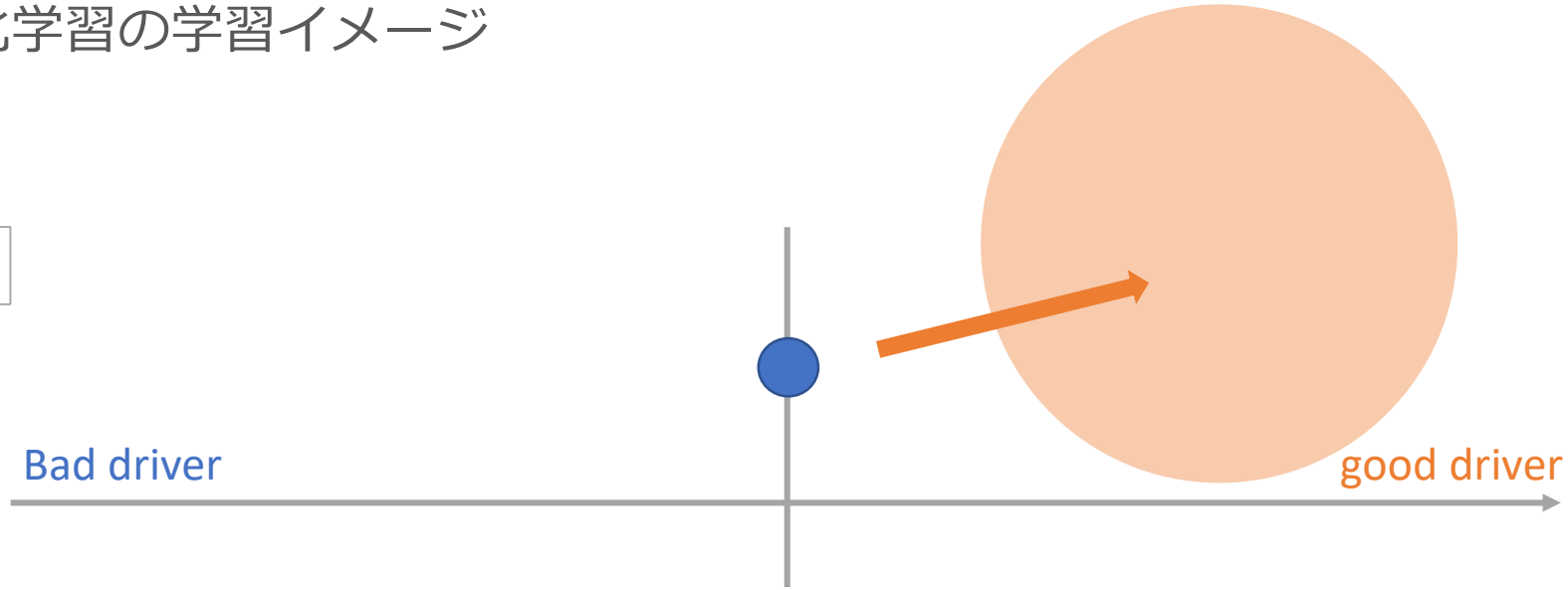
- RLの問題
  - サンプル効率が悪い
    - 初期がランダム探索
  - 膨大な学習時間
    - 10スレッド, 12日間で得られた1000万ステップ
    - (CIRL: 14h + 30万ステップ)

Task	Training conditions				New town				New weather				New town/weather			
	MP	IL	RL	CIRL	MP	IL	RL	CIRL	MP	IL	RL	CIRL	MP	IL	RL	CIRL
Straight	98	95	89	**98**	92	97	74	**100**	100	98	86	**100**	50	80	68	**98**
One turn	82	89	34	**97**	61	59	12	**71**	**95**	90	16	94	50	48	20	**82**
Navigation	80	86	14	**93**	24	40	3	**53**	**94**	84	2	86	47	44	6	**68**
Nav. dynamic	77	**83**	7	82	24	38	2	**41**	**89**	82	2	80	44	42	4	**62**

比較　RL

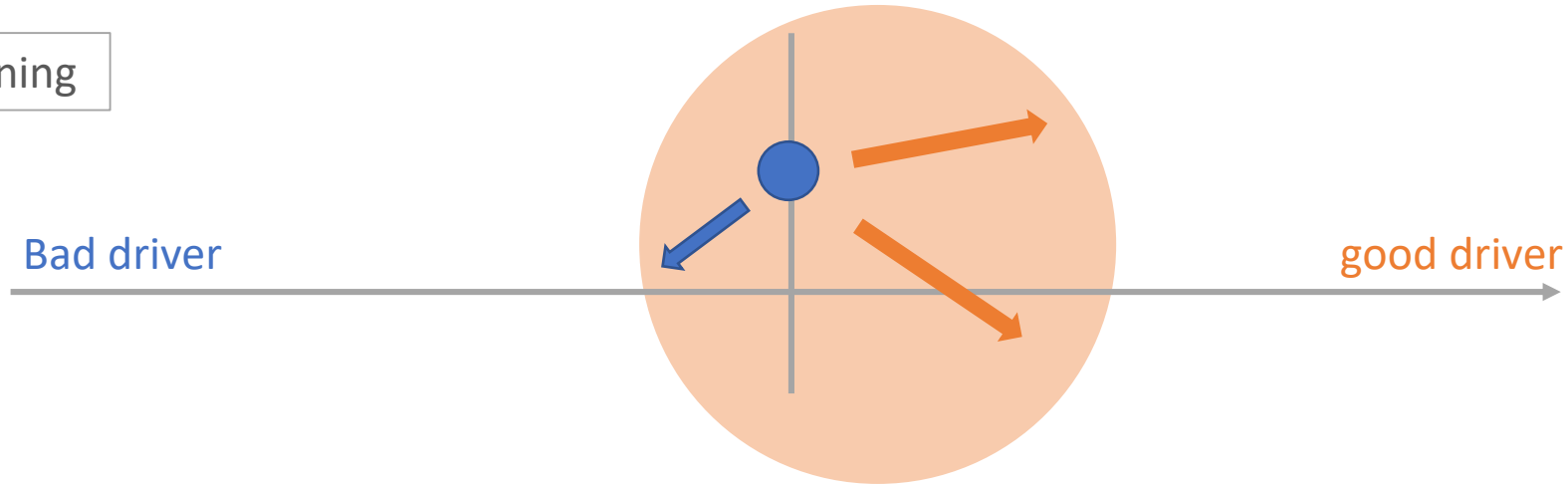
■ 模倣学習と強化学習の学習イメージ

Imitation learning



beginner

Reinforcement learning



比較　IL

■ 定性的評価

Fig. 7. Visualization comparisons between the imitation learning baseline [15] and our CIRL model. We illustrate some driving cases for straight and one-turn tasks, and show the IL baseline fails with some types of infractions (e.g. collision with static object, more than 30% overlap with Sidewalk, in opposite lane) while our CIRL successfully completes the goal-oriented tasks. For each case, two consecutive frames are shown.

比較　MP

- 結果
  - 新天候セットに対して、CIRL より高精度
  - テストデータが学習データに似通っていた

Task	Training conditions				New town				New weather				New town/weather			
	MP	IL	RL	CIRL	MP	IL	RL	CIRL	MP	IL	RL	CIRL	MP	IL	RL	CIRL
Straight	98	95	89	**98**	92	97	74	**100**	100	98	86	**100**	50	80	68	**98**
One turn	82	89	34	**97**	61	59	12	**71**	**95**	90	16	94	50	48	20	**82**
Navigation	80	86	14	**93**	24	40	3	**53**	**94**	84	2	86	47	44	6	**68**
Nav. dynamic	77	**83**	7	82	24	38	2	**41**	**89**	82	2	80	44	42	4	**62**

汎化能力

汎化能力

- ほとんどの条件で CIRL が高い成功率
- 特に new town に対して高い成功率
  - 高い汎化能力の証明
- 新天候は他に劣る
  - 学習データセットの天候に関わるため、汎化能力が低いことにはならない

Task	Training conditions				New town				New weather				New town/weather			
	MP	IL	RL	CIRL	MP	IL	RL	CIRL	MP	IL	RL	CIRL	MP	IL	RL	CIRL
Straight	98	95	89	**98**	92	97	74	**100**	100	98	86	**100**	50	80	68	**98**
One turn	82	89	34	**97**	61	59	12	**71**	**95**	90	16	94	50	48	20	**82**
Navigation	80	86	14	**93**	24	40	3	**53**	**94**	84	2	86	47	44	6	**68**
Nav. dynamic	77	**83**	7	82	24	38	2	**41**	**89**	82	2	80	44	42	4	**62**

Table 2. The percentage (%) of successfully completed episodes of our CIRL on four new settings for further evaluating generalization.

Task	New town/path2	New town/weather2	New path	New weather2
Navigation	50	58	95	87
Nav. dynamic	38	47	87	86

Table 3. The percentage (%) of successfully completed episodes of our CIRL under different weather conditions for the navigation tasks in training town and new town.

Navigation task	CloudyNoon	MidRainyNoon	CloudySunset	WetCloudySunset	HardRainSunset
CIRL (Town 1)	92	96	96	64	56
CIRL (New Town)	95	52	85	90	5

HardRainSunset

実利用への適用

実利用への適用

- リアルシーンでのテスト
  - Comma.ai データセット

- 比較
  - Comma.ai のみでの学習よりも
    CARLA 学習後、Comma.ai チューニング が高精度

Table 5. Results on comma.ai dataset in terms of mean absolute error (MAE).

Model	PilotNet [8]	CIRL (CARLA)	CIRL from scrach	CIRL finetuning
Steer-angle MAE	1.208	2.939	1.186	**1.168**

まとめ

- 模倣学習と強化学習のメリットを組み合わせた
  - 模倣学習
    - 学習の高速収束
  - 強化学習
    - 幅広い探索
- 個別のモデルよりも高い汎化性能
- 手動のルールベース手法よりも高い汎化性能
  - Semantic Segmentation に基づいた制御