

강화학습에 의한 End-to-end 자율주행 TCP 방법의 성능 향상

Enhancing Trajectory-guided Control Prediction of End-to-end Autonomous Driving by Reinforcement Learning

조 기 환¹, 박 태 형^{2*}

(Gi-Hwan Jo¹ and Tae-Hyoung Park^{2,*})

¹Department of Control and Robot Engineering, Chungbuk National University

²Department of Intelligent Systems and Robotics, Chungbuk National University

Abstract: End-to-end autonomous driving models often exhibit limited scalability because they rely on expert demonstration data. Imitation learning techniques such as Trajectory-guided Control Prediction (TCP) face adaptability limitations in varied environments, necessitating extensive data collection and retraining for novel driving scenarios, including static obstacle avoidance. To address this issue, we propose a hybrid framework integrating a pre-trained TCP model with a reinforcement learning (RL) module. The RL agent leverages TCP's latent features and trajectory predictions as its state. By interacting with a simulated environment and optimizing a task-specific reward function, the agent acquires an obstacle avoidance policy, eliminating the need for additional expert data. Experiments in the MORAI simulator demonstrate that our approach reduces vehicle collisions by 53.9% and pedestrian collisions by 45.9% compared to the baseline TCP model, while achieving higher route completion. These findings suggest that augmenting TCP with RL enhances system robustness and enables progressive expansion of driving capabilities, particularly for obstacle avoidance, without additional expert data. While this work focuses on static obstacle avoidance, the proposed TCP with RL framework can be adapted to other driving scenarios by modifying the reward and simulation setup without collecting new expert demonstrations.

Keywords: end-to-end, autonomous driving, reinforcement learning, trajectory-guided control prediction, obstacle avoidance

I. 서론

최근 자율주행 기술은 교통 안전성 향상, 이동 효율성 증대, 경제적 비용 절감과 같은 잠재적 이점으로 인해 학계와 산업계의 주목을 받고 있다. 자율주행 제어 아키텍처는 크게 모듈형 접근법 [1–4]과 E2E(End-to-End) 접근법 [5–10]으로 구분된다. 모듈형 접근법은 인지, 경로 계획, 제어 단계로 세분화되어 각 모듈이 독립적으로 최적화되며, 단계별 문제 진단이 용이하다. 반면, E2E 접근법은 카메라나 LiDAR와 같은 센서 데이터를 직접 입력으로 받아 중간 처리 과정 없이 차량 제어 명령을 출력함으로써 시스템 전반의 통합적 최적화를 목표로 한다.

E2E 자율주행 분야에서 모방학습 기반 접근법은 전문가의 주행 데이터를 활용하여 효과적인 주행 정책을 학습할 수 있다는 장점을 가진다. 특히, TCP(Trajectory-guided Control Prediction) [9]는 단일 카메라 기반의 E2E 자율주행 알고리즘으로, 전방 카메라 이미지, 차량 상태, 그리고 목표점 정보를 입력으로 받아 차량의 제어 명령을 직접 출력한다. TCP는 경로 예측과 제어 예측을 통합하여 수행하며, 목표점 정보를 활용한 통합적 예측 방식으로 기존 모방학습 기반 E2E 모델 대비 경로 완주율을 향상시켰다. 또한 단일 카메라 기반 구조로 학습 및 추론 효율성이 높아 실용적 자율주행 시스템 구현에 적합하다는 장점을 가진다.

그러나 TCP 기반 자율주행 시스템은 성능을 향상시키기 위해서는 전문가 데이터를 추가로 수집하고 재학습해야 하는 근본적인 한계가 있다. 특히, 정적 장애물 회피와 같은 기능을 추가하려면 갓길 주차, 도로 위 보행자 등 장애물 상황에서의 전문가 주행 데이터가 필요하며, 이러한 데이터 생성은 상당한 시간과 비용을 요구한다.

이러한 배경에서 강화학습(reinforcement learning)은 환경과의 상호작용을 통해 스스로 학습할 수 있다는 특성으로 인해 주목받고 있다. 강화학습은 별도의 전문가 데이터 없이도 보상 함수 설계를 통해 새로운 기능을 학습할 수 있으며, 시행착오를 통해 최적의 정책을 탐색한다[11–15]. 따라서 기존 TCP 시스템에 강화학습 모듈을 통합한다면, 전문가 데이터의 추가 수집 없이도 정적 장애물 회피와 같은 새로운 기능을 확장할 수 있다.

본 논문은 이러한 문제를 바탕으로 TCP 기반 자율주행 시스템에 강화학습 모듈을 융합하여 전문가 데이터 없이 정적 장애물 회피 성능을 향상시키는 하이브리드 접근법을 제안한다. 하이브리드 접근법은 처음부터 주행 정책을 학습해야 하는 순수 강화학습 방식과 다르게, 사전 학습된 TCP 모델을 통해 기본적인 차선 유지, 경로 추종 능력과 이미지 특징 추출 능력을 확보한다. 이를 바탕으로 강화학습 모듈은 정적 장애물 회피와 같은 특정 정책 학습에만 집중할 수 있다.

* Corresponding Author

Manuscript received August 5, 2025; revised September 2, 2025; accepted November 28, 2025

조기환: 충북대학교 제어로봇공학 대학원생(ghjo@cbnu.ac.kr, ORCID[®] 0000-0002-7269-1635)

박태형: 충북대학교 지능로봇공학 교수(tachpark@cbnu.ac.kr, ORCID[®] 0000-0002-3695-344x)

※ 본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업임 (IITP-2026-RS-2020-II201462).

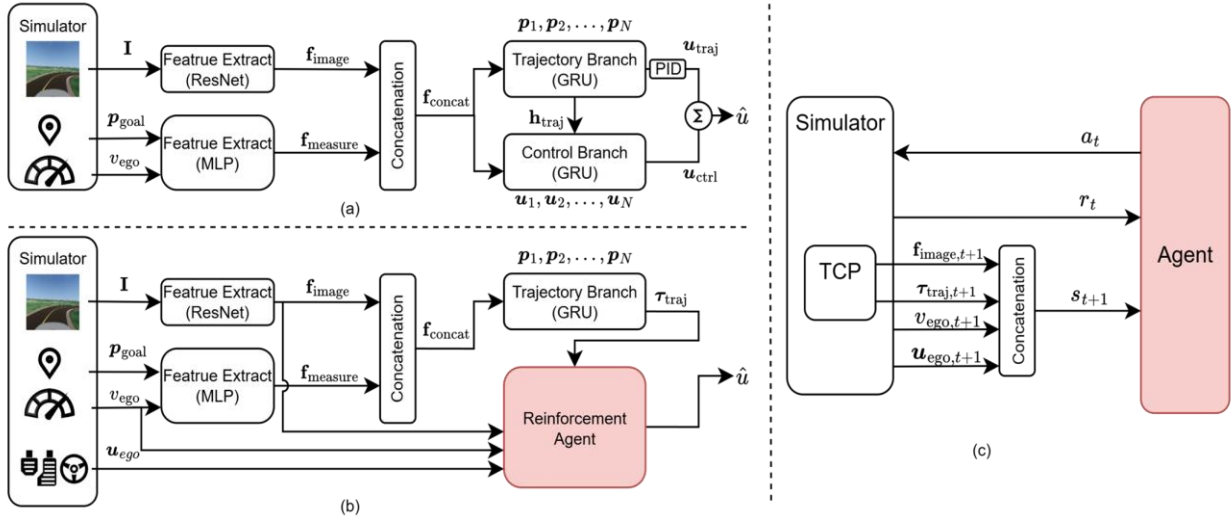


그림 1. (a) 기존 TCP와 (b) 제안하는 Enhanced TCP의 네트워크 구조 비교 및 (c) 강화학습 프레임워크.

Fig. 1. Network structure comparison for (a) conventional TCP, (b) proposed Enhanced TCP, and (c) the reinforcement learning framework.

제안 방법은 사전 학습된 TCP 모델에서 추출한 이미지 특징 벡터와 예측 경로, 그리고 차량의 속도와 제어 출력을 활용하여 강화학습 알고리즘으로 장애물 회피 정책을 학습하고, 추가적인 전문가 데이터 없이 차량제어가 가능한 보상 함수를 설계한다. MORAI 시뮬레이터를 활용한 실험 결과, 제안 모델은 기존 TCP 대비 정적 장애물이 있는 환경에서 충돌률을 낮추고 경로 완주율을 높임으로써 장애물 회피성능의 향상을 입증한다.

본 논문은 다음과 같이 구성된다. 2장에서는 TCP 알고리즘의 구조와 동작 원리를 설명하고, 3장에서는 제안 모델의 구조와 강화학습 알고리즘에 대해 설명한다. 4장에서는 실험 설정과 결과를 분석하며, 5장에서는 연구의 한계와 향후 방향을 논의한다.

II. TRAJECTORY-GUIDED CONTROL PREDICTION

TCP는 단일 카메라 기반 E2E 자율주행 알고리즘으로, 전방 카메라 이미지, 차량 속도, 그리고 목표점 정보를 입력으로 받아 차량의 예측 경로와 제어 명령을 출력한다. 그림 1(a)는 TCP의 시스템 구조를 보여준다. 주요 특징은 경로 예측과 제어 예측을 분리하지 않고, 융합된 특징 벡터를 활용하여 두 과정을 통합적으로 수행한다는 점이다. TCP는 경로 예측과 제어 예측 모듈의 결과를 융합하여 기존의 단순한 행동 예측 방식 대비 경로 완주율을 향상시켰다.

1. 네트워크 구조 및 입출력

TCP 네트워크의 입력은 전방 카메라 이미지 $I \in \mathbb{R}^{H \times W \times 3}$, 현재 차량 속도 v_{ego} , 그리고 목표점 좌표 $p_{goal} = [p_x, p_y]$ 이다. 네트워크는 향후 N 개 시점의 예측 경로인 $\tau_{traj} = [p_1, p_2, \dots, p_N]$ 와 가속도 및 조향각으로 구성된 예측 제어명령 $\tau_{ctrl} = [u_1, u_2, \dots, u_N]$ 을 산출한다. 최종적으로 이 두 예측 값을 이용하여 $\hat{u} = [\hat{u}_{acc}, \hat{u}_{steer}]$ 를 출력한다.

2. 특징 추출

TCP는 입력 데이터를 두 가지 경로로 처리한다. 먼저, 식 1과 같이 전방 카메라 이미지 I 를 CNN (Convolutional Neural

Network) 기반의 ResNet [16]에 입력하여 이미지 특징 벡터 $f_{image} \in \mathbb{R}^{d_i \times 1}$ 로 인코딩한다.

$$f_{image} = \text{CNN}(I) \quad (1)$$

동시에, 차량 속도 v_{ego} 와 목표점 좌표 p_{goal} 를 별도의 MLP (Multi-Layer Perceptron) 기반 임베딩 네트워크에 입력하여, 식 2와 같이 측정 특징 벡터 $f_{measure} \in \mathbb{R}^{d_m \times 1}$ 을 생성한다.

$$f_{measure} = \text{MLP}([v_{ego}, p_x, p_y]) \quad (2)$$

각 특징벡터는 concatenation을 통해 통합 특징 벡터 $f_{concat} \in \mathbb{R}^{(d_i + d_m) \times 1}$ 로 융합된다:

$$f_{concat} = [f_{image}, f_{measure}] \quad (3)$$

이 통합 특징 벡터는 경로 및 제어 예측 단계로 전달된다.

3. 경로 및 제어 예측

TCP는 경로 예측 모듈과 제어 예측 모듈로 구성된 병렬 아키텍처를 통해 주행 정책을 결정한다. 경로 예측 모듈은 GRU (Gated Recurrent Unit) [17]를 활용하여 시간 시퀀스 데이터를 처리함으로써 예측 경로 τ_{traj} 를 생성한다. 이 과정에서 GRU는 은닉 상태 $h_{traj,t}$ 를 업데이트하며, 별도의 MLP네트워크를 이용하여 예측 경로의 위치 p_{t+1} 를 예측한다:

$$h_{traj,t+1} = \text{GRU}_{traj}(f_{concat}, h_{traj,t}) \quad (4)$$

$$p_{t+1} = \text{MLP}(h_{traj,t}) \quad (5)$$

생성된 예측 경로는 PID 제어기를 통해 안정적인 추종을 위한 제어 신호 u_{traj} 로 변환된다.

제어 예측 모듈은 GRU 기반 다중 단계 제어 예측(multi-step control prediction)을 도입하였다. 경로 예측 모듈에서 출력된 은닉 상태 h_{traj} 와 통합 특징벡터 f_{concat} 을 결합하여 일련의 제어 명령 τ_{ctrl} 을 예측한다.

두 모듈의 출력은 가중치 조절 메커니즘을 통해 최종 제어 신호 $\hat{u}$ 로 통합된다:

$$\hat{\mathbf{u}} = \alpha \mathbf{u}_{\text{ctrl}} + (1 - \alpha) \mathbf{u}_{\text{traj}} \quad (6)$$

여기서 $\mathbf{u}_{\text{ctrl}}$ 는 τ_{ctrl} 의 첫번째 제어 명령이며, 가중치 $\alpha \in [0,1]$ 는 주행 상황에 따라 동적으로 조정된다. 주행 상황은 이전 조향각 절대값 20개 중 절반 이상이 0.1을 초과하면 회전 상황으로, 그 외는 직진상황으로 판단한다. 회전 상황에서는 α 을 높여 $\mathbf{u}_{\text{ctrl}}$ 의 비율을 높이고, 직진 상황에서는 반대로 $\mathbf{u}_{\text{traj}}$ 의 비율을 높인다.

4. TCP의 장점과 한계

TCP는 목표점 정보를 활용한 통합적 예측 방식으로 카메라를 사용하는 기존 E2E 모델 대비 경로 완주율을 향상시켰다. 그러나 모방학습에 의존하는 근본적 한계로 인해 성능을 향상시키기 위해서는 해당 상황에 맞는 전문가 데이터를 추가로 수집하고 재학습해야 한다. 특히, 정적 장애물 회피와 같은 경우 다양한 위험 상황에서의 전문가 데이터 수집이 비용과 시간이 많이 소요된다는 문제가 있다.

III. 제안 방법

1. 시스템 아키텍처

본 논문에서 제안하는 시스템은 기존의 TCP에 강화학습 모듈을 융합하여, 전문가 데이터의 추가 수집 없이 정적 장애물 회피라는 새로운 기능을 추가하는 것을 목표로 한다. 그림 1(b)는 전체 구조를 도식화한 것이다.

제안하는 시스템은 두 가지 핵심 모듈로 구성된다. 첫 번째는 사전 학습된 TCP 모듈로, 기존 구조를 유지하되, 이미지 특징벡터 $\mathbf{f}_{\text{image}}$ 와 예측 경로 τ_{traj} 를 강화학습 모듈에 전달한다. 두 번째는 강화학습 모듈로, TCP로부터 전달받은 이미지 특징벡터와 예측 경로, 그리고 시뮬레이터에서 제공하는 현재 차량 속도 v_{ego} 와 제어 입력 값 $\mathbf{u}_{\text{ego}}$ 를 상태(state)로 정의한다. 이를 바탕으로 정적 장애물 회피를 위한 최적의 제어 명령 $\hat{\mathbf{u}}$ 을 출력한다. 이 모듈은 환경과의 상호작용을 통해 스스로 장애물 회피 전략을 학습한다.

이러한 구조를 통해, 기존 TCP 모델이 학습하지 않은 장애물 회피 상황에서도 전문가 데이터 없이 안전한 주행이 가능하다. 강화학습 모듈은 환경과의 상호작용을 통해 스스로 장애물 회피 전략을 학습하므로, 다양한 장애물 상황에 대한 별도의 전문가 데이터 수집이 불필요하다.

2. 강화학습 모델 설계

강화학습 프레임워크는 그림 1(c) 나타나 있다. 에이전트(agent), 환경(environment), 그리고 상태(state)-행동(action)-보상(reward) 순환 구조로 구성된다. 에이전트는 환경으로부터 현재 상태 s_t 를 관측하고 정책 π 에 따라 행동 a_t 를 선택한다. 환경은 에이전트의 행동에 따라 다음 상태 s_{t+1} 로 전이되며, 보상 r_t 을 에이전트에게 제공한다. 이러한 상호작용을 통해 에이전트는 누적 보상을 최대화하는 정책을 학습한다.

2.1 상태 공간 (state space)

상태 s_t 는 에이전트가 주어진 시점 t 에서 인지할 수 있는 모든 정보를 포함한다. 본 연구에서는 다음과 같이 정의한다

$$\mathbf{s}_t = [\mathbf{f}_{\text{image},t}, \tau_{\text{traj},t}, v_{\text{ego},t}, \mathbf{u}_{\text{ego},t}] \in \mathcal{S} \quad (7)$$

여기서 $\mathbf{f}_{\text{image},t}$ 와 $\tau_{\text{traj},t}$ 는 t 시점에서 TCP 네트워크에서

추출된 이미지 특징 벡터와 예측 경로이다. v_t 와 $\mathbf{u}_t$ 는 차량속도와 제어 명령 값이다.

2.2 행동공간(action space)

행동 a_t 는 에이전트가 시뮬레이터에 입력하는 제어 명령으로 구성되고, 다음과 같이 정의한다.

$$\mathbf{a}_t = [u_{\text{acc},t}, u_{\text{steer},t}] \in \mathcal{A} \quad (8)$$

여기서 $u_{\text{acc},t} \in [-1,1]$ 는 가속/감속(양수: 가속, 음수: 감속 또는 브레이크), $u_{\text{steer},t} \in [-1,1]$ 는 조향각을 의미한다.

2.3 보상 함수(reward function)

강화학습 알고리즘 학습을 위한 보상 함수 r 은 객체 회피와 주행 효율성, 그리고 제어 신호의 안정성을 고려하여 설계되었다. 보상 함수는 다음과 같은 항목들로 구성된다.

충돌 페널티 $r_{\text{collision}}$ 는 차량이 객체와 충돌할 경우 페널티를 부여하여 충돌을 회피하도록 유도한다.

$$r_{\text{collision}} = \begin{cases} -1, & \text{if collision occurs} \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

경로 횡방향 오차 페널티 r_{lateral} 는 차량의 현재 위치와 전역 경로 사이의 횡방향 오차에 비례하는 음의 보상을 부여하여 경로를 벗어나지 않도록 유도한다. 오차는 L_2 norm을 사용하여 계산된다.

$$r_{\text{lateral}} = \|\mathbf{p}_{\text{ego}} - \mathbf{p}_{\text{path_closest}}\| \quad (10)$$

여기서 $\mathbf{p}_{\text{ego}}$ 은 차량의 현재 위치, $\mathbf{p}_{\text{path_closest}}$ 는 전역 경로 상 차량에 가장 가까운 지점이다.

각도 오차 페널티 r_{angle} 는 차량의 heading 각도와 전역 경로의 진행 방향 각도 사이의 오차에 비례하는 음의 보상을 부여한다.

$$r_{\text{angle}} = |\theta_{\text{ego}} - \theta_{\text{path}}| \quad (11)$$

여기서 θ_{ego} 는 차량의 heading 각도, θ_{path} 는 전역 경로의 접선 각도이다.

목표 속도 달성 보상 r_{speed} 는 차량의 현재 속도가 목표 속도에 가까워질수록 양의 보상을 부여한다. 목표 속도와 정확히 일치할 때 최대 보상 1을, 속도 차이가 커질수록 0에 가까워지는 보상을 부여한다.

$$r_{\text{speed}} = 1 - \frac{|v_{\text{ego}} - v_{\text{target}}|}{v_{\text{target}}} \quad (12)$$

여기서 v_{ego} 는 현재 차량 속도, v_{target} 는 목표 속도(30km/h)이다.

전체 보상 함수는 다음과 같이 정의된다.

$$r = w_1 r_{\text{collision}} + w_2 r_{\text{lateral}} + w_3 r_{\text{angle}} + w_4 r_{\text{speed}} \quad (13)$$

w_1, w_2, w_3, w_4 는 각 보상 별 가중치이며, 우선순위에 따라 충돌, 횡방향 오차 및 각도 오차, 속도 순으로 설정하여, 안전성을 최우선으로 고려하면서도 경로 유지 및 효율적인 주행을 유도한다.

3. 강화학습 알고리즘

강화학습 알고리즘은 주어진 환경에서 할인된 누적 보상을 최대화하는 최적 정책을 학습한다는 공통 목표를 가진다.

$$J(\pi) = E_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \quad (14)$$

여기서 π 는 정책, τ 는 에피소드, $\gamma \in [0,1)$ 는 할인 인자, $r(s_t, a_t)$ 는 상태-행동 쌍에 대한 보상이다. 본 연구에서는 자율주행 분야에서 널리 사용되는 세 가지 대표적인 알고리즘 DDQN (Double Deep Q-Network) [18]과 PPO (Proximal Policy Optimization) [19] 그리고 SAC (Soft Actor-Critic) [20]를 선택하여 강화학습 정책을 학습하였다.

DDQN은 전통적인 DQN의 Q 값 과대평가 문제를 해결하기 위해 제안된 가치 기반 강화학습 알고리즘이다. DDQN은 행동 선택과 Q 값 평가를 분리함으로써, 타깃 값 계산 과정에서 발생하는 편향을 완화한다. 이는 메인 네트워크로 최적 행동을 결정하고, 타깃 네트워크로 해당 행동의 Q 값을 평가함으로써 이루어진다. 또한 경험 재현(experience replay)과 타깃 네트워크를 활용하여 학습 안정성과 샘플 효율성을 향상시킨다. 그러나 DDQN은 이산 행동 공간에 기반하므로, 자율주행과 같은 연속 제어 과제에는 이산 행동으로의 변환이 필요하다. 이로 인해 제어 정밀도가 저하될 수 있다.

PPO는 정책 경사 방법의 일종으로 정책 업데이트 시 정책 변화의 크기를 제한함으로써 과도한 변화를 방지하고, 안정적인 학습을 가능하게 한다. 또한 연속 행동 공간을 직접 처리할 수 있어 조향 및 가속 제어에 적합하다. 그러나 PPO는 온폴리시(on-policy) 알고리즘으로, 매 학습 스텝마다 새로운 경험을 수집해야 하므로 샘플 효율성이 낮아, 복잡한 장애물 회피 과제에서는 많은 시간이 소요될 수 있다.

SAC은 최대 엔트로피 프레임워크를 기반으로 한 오프폴리시(off-policy) 액터-크리틱 알고리즘으로, 보상 최대화와 함께 정책의 엔트로피도 동시에 극대화한다. 이는 효율적인 탐색을 유도하며, 환경 변화에 대한 견고성을 제공한다. SAC는 연속 행동 공간에서의 직접적인 제어가 가능하며, 오프폴리시 학습 방식으로 인해 높은 샘플 효율성을 갖는다. 또한 자동 엔트로피 튜닝(automatic entropy tuning)을 통해 학습 안정성과 성능을 향상시킨다. 이러한 특성은 자율주행에서의 정밀한 제어, 효율적인 학습에서 유리하다.

본 연구에서는 자율주행 환경의 연속적 행동 공간, 높은 샘플 효율성 요구, 그리고 안정적 학습의 필요성을 종합적으로 고려하여 SAC 알고리즘을 채택하였다.

IV. 실험

1. 실험 환경 구성

본 논문에서 제안하는 방법의 성능을 평가하기 위하여 MORAI 시뮬레이터를 사용하여 실험을 진행하였다. 모든 실험은 단일 NVIDIA RTX 3090 GPU 환경에서 수행되었다. 실험에는 화성 자동차안전연구원의 K-city 환경을 사용하였으며, 도심 환경에 전역경로를 설정하여 실험을 진행하였다. 실험은 갓길에 주차된 차량, 도로 위에 정지한 차량, 그리고 도로 위 보행자와 같은 정적 장애물을 포함하는 서로 다른 28개의 시나리오를 구성하여 실험을 진행하였다.

평가 항목은 경로 완주율(route complete)과, 충돌률(collision rate) 두 가지로 구성하였다. 경로 완주율은 시나리오 목표경

로의 출발지로부터 목적지까지 완주 비율, 충돌률은 차량과 보행자 각각 전체 장애물 수 대비 충돌한 장애물 수의 비율을 나타낸다.

$$\text{경로 완주율(\%)} = \frac{\text{진행한 경로의 길이(m)}}{\text{전체 경로의 길이(m)}} \times 100 \quad (15)$$

$$\text{충돌률(\%)} = \frac{\text{충돌한 장애물 수}}{\text{전체 장애물 수}} \times 100 \quad (16)$$

TCP모델 학습에 사용된 전문가 데이터는 MORAI 시뮬레이터의 자율주행 기능을 이용하여 전방 이미지, 차량위치, 차량 제어 값을 수집하였다. K-city의 도심환경과 어린이 보호구역에서 총 200,000 프레임의 전문가 데이터를 취득하였으며, 이를 이용해 TCP모델을 20epoch 동안 학습시켰다.

강화학습 보상함수의 가중치는 실험적으로 결정하였다.

$$w_1 = -50, w_2 = -0.5, w_3 = -0.5, w_4 = 1$$

2. 실험 결과

표 1은 각 강화학습 알고리즘을 100만 스텝 동안 학습시켰을 때의 시간 및 자원 소모량을 비교한 결과이다. 단일 연산 시간과 GPU 메모리 사용량은 DDQN, PPO, SAC 순으로 효율적이었으나, 학습 시간은 SAC, DDQN, PPO 순으로 짧은 것으로 나타났다. 학습 시간과 연산 시간이 비례하지 않는 주된 원인은 시뮬레이터의 병목현상 때문이다. PPO 알고리즘은 시뮬레이터 환경 초기화 요청이 가장 많아 학습 시간이 가장 길었다.

그림 2는 DDQN, PPO, SAC 세 알고리즘의 학습 보상 곡선을 보여준다. 진한 선은 보상함수를 지수 가중평균한 값을 나타낸다. 선 주위의 연한색의 범위는 주변 10개 값들의 표준편차이다. DDQN의 경우, 약 50만 스텝까지 보상이 증가

표 1. 세 가지 강화학습 알고리즘의 학습 시간, GPU 메모리, 연산 시간 비교

Table 1. Comparison of learning time, GPU memory, and computation time for three reinforcement learning algorithms.

Method	Learning Step	Learning Time(h)	GPU Memory (MB)	Computation Time(ms)
DDQN	1,000,000	37.6	4.4	0.21
PPO	1,000,000	39.9	4.5	0.28
SAC	1,000,000	31.2	11.4	0.32

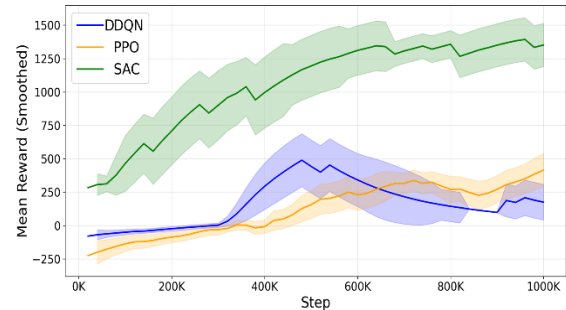


그림 2. 강화학습 알고리즘 별 보상학습 그래프.

Fig. 2. Reinforcement learning algorithm-specific reward learning graph.

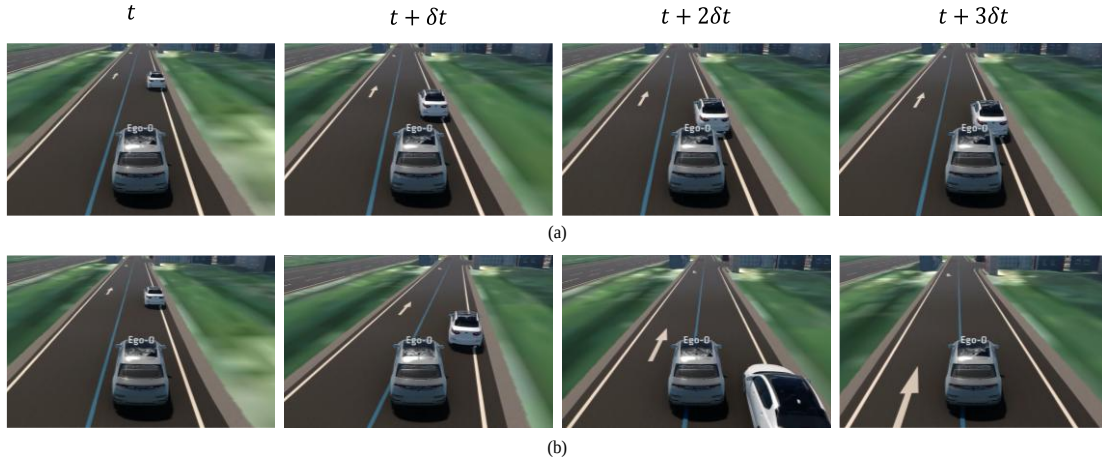


그림 3. MORAI 시뮬레이터 환경에서 정지 객체에 대한 (a) 기존 TCP와 (b) 제안된 방법의 회피 기동 비교.

Fig. 3. Comparison of avoidance maneuvers for static objects between (a) conventional TCP and (b) proposed method in MORAI simulator environment.

하는 경향을 보였으나, 이후 학습이 불안정해지며 성능이 저하되는 현상이 관찰되었다. PPO 는 보상이 점진적으로 상승하고, 표준편차가 작아 안정적인 학습 패턴을 보여주었다. 하지만 상대적으로 느린 수렴 속도를 보여 제한 스텝 안에서 충분히 학습하지 못하였다. SAC는 초기 학습 단계에서 가장 빠른 학습 속도를 보였으며, 약 60만 스텝에서 수렴하였다.

그림 3은 제안하는 알고리즘이 갓길 주차된 차량을 회피하는 시나리오를 묘사하며, 시간에 따른 차량의 거동을 보여준다. 그림 3(a)는 기존 TCP 방법의 주행 결과로, 차량을 회피하지 못하고 전방 차량과 충돌하는 상황을 보여준다. 반면, 그림 3(b)는 본 연구에서 제안하는 방법의 주행결과로, 안전하게 장애물을 회피하는 성공했음을 보여준다. 회피 기동에 대해 정량적으로 분석하면, 기존 TCP는 평균 횡방향 오차 0.1m, 평균 회피시간 1.0초로 차선 이탈이 작았으나 충돌이 빈번하게 발생했다. 제안 방법은 평균 횡방향 오차 1.1m, 평균 회피시간 1.3초로 더 큰 안전 거리를 확보하며 회피 기동을 수행하였다.

표 2는 정적 장애물(주차 차량, 보행자)이 배치된 환경에서 TCP, TCP+SAC, 그리고 비교 모델들의 성능을 비교한 결과이다. 각 모델은 28개의 서로 다른 시나리오에서 10회씩 반복 주행하였으며, 차량 충돌률, 보행자 충돌률, 그리고 주행 성공률의 평균을 나타낸다. TCP 방법은 차량과 보행자 충돌률이 각각 87.3%, 66.7%로 충돌률이 가장 높았다. 또한 경로 완주율도 27.1%로 가장 낮은 성능을 보였다. 반면, 제안한 TCP+SAC 방법은 기존 방법 대비 경로 완주율이 50.4%증가했으며, 차량과 보행자 충돌률은 각각 53.9%, 45.9% 감소하였다. 실험에서 TCP+SAC 모델이 가장 우수한 성능을 보인 이유는 경험재생과 엔트로피 최대화를 통한 탐험 성능 향상으로 파악된다. TCP+PPO는 경험 재현이 없어 각 샘플의 효율이 떨어진다. 최종적으로 제한된 스텝안에 충분히 학습하지 못해 강화학습을 융합한 방법 중 가장 낮은 성능을 보였다. TCP+DQN은 경험 재현이 있지만, 학습이 진행됨에 따라 탐사 비율을 늘리는 방법을 사용한다. 이 때문에 충분히 탐험하지 않았을 때 탐사를 진행하여, 중간 이후부터 오히려 성능이 감소하는 것을 확인할 수 있었다.

표 2. 기존 방법과 제안 방법의 경로 완주율, 충돌률 성능 비교.

Table 2. Compare of route complete and collision rate performance between traditional and proposed methods.

Method	Route Complete (%)	Collision Rate (%)	
		Vehicle	Pedestrian
TCP	27.1	87.3	66.7
TCP+DDQN	71.9	54.2	27.1
TCP+PPO	40.8	55.0	41.2
TCP+SAC	77.5	33.4	20.8

V. 결론

본 논문에서는 모방학습 기반 E2E 자율주행 모델인 TCP에 강화학습 모듈(SAC)을 결합한 하이브리드 프레임워크를 제안하였다. 모방학습 기반 E2E 모델은 전문가 데이터에 의존하기 때문에, 새로운 주행 기능을 추가하거나 학습되지 않은 상황에 대응할 때 추가 데이터 수집 및 재학습이 필요하다는 한계를 가진다. 제안 방법은 이러한 데이터 의존성 문제를 강화학습을 통해 완화하였으며, TCP의 이미지 특징 벡터와 예측 경로 정보를 활용하여 추가 전문가 데이터 없이 정적 장애물 회피 기능을 학습할 수 있도록 설계되었다.

MORAI 시뮬레이터를 통한 실험에서, TCP+SAC는 기존 TCP 대비 차량 충돌률을 53.9%, 보행자 충돌률을 45.9% 감소시켰다. 이는 강화학습 모듈이 환경과의 상호작용을 통해 스스로 회피 전략을 학습할 수 있음을 보여준다.

본 연구는 모방학습 기반 E2E 모델의 구조적 한계인 전문가 데이터 의존성을 강화학습 결합을 통해 개선하는 방법을 제시하였다. 이러한 접근법은 TCP뿐만 아니라 유사한 구조를 가진 다른 모방학습 기반 E2E 자율주행 모델에도 적용할 수 있을 것으로 기대된다. 다만, 시뮬레이터 환경의 한정된 시나리오에서만 평가되었으므로, 실제 도로 환경에서의 일반화 성능과 실시간 적용 가능성에 대한 추가 검증이 필요하다. 향후 연구에서는 실제 차량 실험, 다양한 주행 시나리오 평가, 강화학습 모듈의 효율적 학습 및 경량화와 관련된 연구를 수행할 계획이다.

REFERENCES

- [1] H.-J. Sim, J. Kim, J. Hong, S.-W. Nam, Y.-H. Kim, J. Heo, H.-C. Hwang, and K.-K. Kim, "Visual navigation in unstructured environments and the development of a multi-purpose mobile robot platform," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 9, pp. 913–923, Sep. 2024.
- [2] M. Testouri, G. Elghazaly, and R. Frank, "RoboCar: A rapidly deployable open source platform for autonomous driving research," *IEEE Intelligent Transportation Systems Magazine*, vol. 17, no. 4, pp. 83–95, Jul. 2025.
- [3] M. Heinrich, M. Zipfl, M. Uecker, S. Ochs, M. Gontscharow, T. Fleck, J. Doll, P. Schömer, C. Hubschneider, M. R. Zofka, A. Viehl, and J. M. Zöllner, "CoCar NextGen: A multi-purpose platform for connected autonomous driving research," *Proc. of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems*, Edmonton, Canada, pp. 482–489, Sep. 2024.
- [4] W. Zhang, M. Elmahgiubi, K. Rezaee, B. Khamidehi, H. Mirkhani, F. Arasteh, C. Li, M. A. Kaleem, E. R. Corral-Soto, D. Sharma, and T. Cao, "Analysis of a modular autonomous driving architecture: The top submission to CARLA leaderboard 2.0 challenge," arXiv, Mar. 2024.
- [5] F. Codevilla, M. Muller, A. Lopez, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," *Proc. of the 2018 IEEE International Conference on Robotics and Automation*, Brisbane, Australia, pp. 4693–4700, May. 2018.
- [6] X. Liang, T. Wang, L. Yang, and E. Xing, "CIRL: Controllable imitative reinforcement learning for vision-based self-driving," *Proc. of the European Conference on Computer Vision*, Munich, Germany, pp. 604–620, Sep. 2018.
- [7] F. Codevilla, E. Santana, A. Lopez, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," *Proc. of the 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, pp. 9328–9337, Oct. 2019.
- [8] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," *Proc. of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 7073–7083, Jun. 2021.
- [9] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," *Proc. of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 6119–6132, Jun. 2022.
- [10] J.-H. Lee, J.-Y. Yeon, and T.-H. Park, "End-to-end network for autonomous vehicles based on a bird's eye view perception module," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 31, no. 8, pp. 921–926, Aug. 2025.
- [11] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," *Proc. of the 2021 IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp. 15202–15212, Oct. 2021.
- [12] P. Maramotti, A. P. Capasso, G. Bacchiani, and A. Broggi, "Tackling real-world autonomous driving using deep reinforcement learning," *Proc. of the 2022 IEEE Intelligent Vehicles Symposium*, Aachen, Germany, pp. 1274–1281, Jun. 2022.
- [13] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, Jun. 2022.
- [14] J. Chen, B. Yuan, and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," *Proc. of the 2019 IEEE Intelligent Transportation Systems Conference*, Auckland, New Zealand, pp. 2765–2771, Oct. 2019.
- [15] M. Park, J.-S. Hong, and N.-K. Kwon, "Deep reinforcement learning based on the hindsight experience replay for autonomous driving of mobile robot," *Journal of Institute of Control, Robotics and Systems(in Korean)*, vol. 28, no. 11, pp. 1006–1012, Nov. 2022.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, Jun. 2016.
- [17] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1724–1734, Oct. 2014.
- [18] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," arXiv, Dec. 2015.
- [19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv, Aug. 2017.
- [20] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *Proc. of the 35th International Conference on Machine Learning*, Stockholm, Sweden, pp. 1861–1870, Jul. 2018.



조 기 환

2019년 충북대학교 전자공학부 졸업.
2022년 충북대학교 제어로봇공학전공 석사 졸업. 2022년~현재 충북대학교 대학원 제어로봇공학 박사과정 재학 중.
관심분야는 자율주행, 강화학습.



박 태 형

1988년 서울대 제어계측공학과 졸업.
1990년 동 대학원 석사. 1994년 동 대학원 박사. 1994년~1997년 삼성테크윈(주) 선임연구원 1997년~현재 충북대학교 지능로봇공학과 교수. 2000년~2001년 토론토 대학교 방문교수. 2012년~2017년 대한전기학회 로보틱스 및 자동차 연구회 위원장. 2020년~현재 과기정통부 지정 산업인공지능 연구센터 (Grand ICT) 센터장.
관심분야는 로보틱스, 자율주행, 산업인공지능 등.