The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 24, No. 3, pp.15-20, Jun. 30, 2024. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2024.24.3.15
JIIBC 2024-3-3

셀 분해 알고리즘을 활용한 심층 강화학습 기반 무인 항공기 경로 계획

UAV Path Planning based on Deep Reinforcement Learning using Cell Decomposition Algorithm

김경훈*, 황병선*, 선준호*, 김수현*, 김진영**

Kyoung-Hun Kim*, Byungsun Hwang*, Joonho Seon*,
Soo-Hyun Kim*, Jin-Young Kim**

요 약 무인 항공기의 경로 계획은 고정 및 동적 장애물을 포함하는 복합 환경에서 장애물 충돌을 회피하는 것이 중요하다. RRT나 A^*와 같은 경로 계획 알고리즘은 고정된 장애물 회피를 효과적으로 수행하지만, 고차원 환경일수록 계산 복잡도가 증가하는 한계점을 가진다. 강화학습 기반 알고리즘은 복합적인 환경 반영이 가능하지만, 기존 경로 계획 알고리즘과 같이 고차원 환경일수록 훈련 복잡도가 증가하여 수렴성을 기대하기 힘들다. 본 논문은 셀 분해 알고리즘을 활용한 강화학습 모델을 제안한다. 제안한 모델은 학습 환경을 세부적으로 분해하여 환경의 복잡도를 감소시킨다. 또한, 에이전트의 유효한 행동을 설정하여 장애물 회피 성능을 개선한다. 이를 통해 강화학습의 탐험 문제를 해결하고, 학습의 수렴성을 높인다. 시뮬레이션 결과는 제안된 모델이 일반적인 환경의 강화학습 모델과 비교하여 학습 속도를 개선하고 효율적인 경로를 계획할 수 있음을 보여준다.

Abstract Path planning for unmanned aerial vehicles (UAV) is crucial in avoiding collisions with obstacles in complex environments that include both static and dynamic obstacles. Path planning algorithms like RRT and A^* are effectively handle static obstacle avoidance but have limitations with increasing computational complexity in high-dimensional environments. Reinforcement learning-based algorithms can accommodate complex environments, but like traditional path planning algorithms, they struggle with training complexity and convergence in higher-dimensional environment. In this paper, we proposed a reinforcement learning model utilizing a cell decomposition algorithm. The proposed model reduces the complexity of the environment by decomposing the learning environment in detail, and improves the obstacle avoidance performance by establishing the valid action of the agent. This solves the exploration problem of reinforcement learning and improves the convergence of learning. Simulation results show that the proposed model improves learning speed and efficient path planning compared to reinforcement learning models in general environments.

Key Words : A2C, Cell Decomposition, Path Planning, Reinforcement Learning, UAV

*준회원, 광운대학교 전자융합공학과
**정회원, 광운대학교 전자융합공학과, 교신저자
접수일자 2024년 4월 4일, 수정완료 2024년 5월 4일
게재확정일자 2024년 6월 7일

Received: 4 April, 2024 / Revised: 4 May, 2024 /
Accepted: 7 June, 2024
**Corresponding Author: jinyoung@kw.ac.kr
Dept. of Electronic Convergence Engineering, Kwangwoon University, Korea

Ⅰ. 서 론

무인 항공기는 비용 효율성, 높은 기동성, 비행경로의 자율성으로 인해 군사, 상업 및 민간 분야에서 활용되고 있다[1]. 무인 항공기의 자율적인 경로 계획은 빌딩, 나무, 전선 등과 같은 고정 장애물과 새, 비행물체 등 동적 장애물과 상호작용하는 능력이 요구된다. 따라서 무인 항공기의 안전하고 효율적인 운용을 위해 장애물들과의 충돌을 피할 수 있는 경로 계획 알고리즘이 필수적이다.

기존의 경로 계획 알고리즘인 RRT (rapid-exploring random tree)와 A^*알고리즘은 고정된 장애물이 있는 환경에서 최적 경로를 결정할 수 있지만, 환경 크기에 따라 계산 복잡도가 급격하게 증가하는 한계가 존재한다[2, 3]. 이를 보완하기 위해 제안된 강화학습 기반 경로 계획 알고리즘은 A^* 알고리즘과 비교하였을 때, 최적 경로 탐색 시간은 단축되는 것이 확인되었지만, 장애물 회피 성능이 부족하다[4, 5]. 또한, 심층 Q 네트워크(Deep-Q Network, DQN)를 이용한 경로 계획 학습은 고차원 환경에서 학습에 필요한 메모리의 양이 증가하여 학습의 복잡도가 높아지고, 학습 공간의 탐험 부족으로 인해 학습이 발산할 수 있는 한계가 있다[6].

강화학습을 활용한 무인 항공기 경로 계획은 복합적인 환경에서 최적 경로와 장애물을 회피할 수 있지만, 학습 환경의 복잡도에 따라서 에이전트의 성능이 제한될 수 있다. 본 논문에서는 환경에 셀 분해 알고리즘을 적용한 강화학습 모델을 제안한다. 제안된 모델은 A2C 알고리즘을 사용하며, 고차원 환경을 셀 분해 알고리즘을 통해 학습 환경 복잡도를 낮추어 강화학습의 학습 능력을 개선하고, 신뢰도 높은 장애물 회피 성능을 가진다.

그림 1. 강화학습 알고리즘 모식도.
Fig. 1. Schematic diagram of reinforcement learning algorithm.

본 논문의 구성은 다음과 같다. 2장에서는 A2C와 셀 분해 알고리즘을 설명하고, 3장에서는 셀 분해 알고리즘을 활용한 환경 구성과 강화학습 모델을 정의한다. 4장에서는 제안한 모델의 시뮬레이션 결과를 확인한다.

Ⅱ. A2C 및 셀 분해 알고리즘

1. A2C 강화학습 알고리즘

강화학습은 기계학습의 한 분야로, 그림 1과 같이 에이전트가 환경과 상호작용하여 보상을 최대화하는 방식으로 학습한다. 에이전트는 마르코프 결정 과정을 통해 주어진 상태(state)에서 행동(action)을 결정하고, 행동의 결과로 환경으로부터 보상(reward)을 받는다. 에이전트는 환경의 탐험으로 누적된 보상을 최대화하는 정책(policy)을 학습한다.

A2C(Advantage Actor Critic)는 그림 2와 같이 액터 네트워크와 크리틱 네트워크로 구성된 강화학습 알고리즘이다. 액터 네트워크는 학습된 정책을 기반으로 에이전트의 가능한 행동을 결정한다. 비평가 네트워크는 가치 기반 네트워크로 에이전트 행동의 상태 가치를 평가한다. 두 네트워크는 수식 1을 최대화하기 위해 업데이트된다.

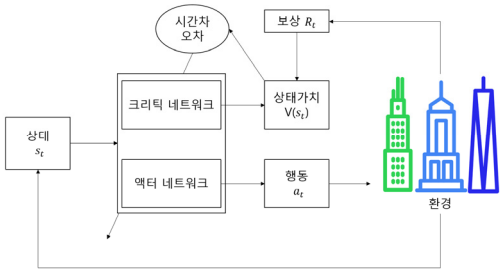

그림 2. A2C 알고리즘.
Fig. 2. Schematic diagram of A2C.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[A(s_t \mid a_t)\nabla_\theta \pi_\theta \log(a_t \mid s_t)]. \quad (1)$$

수식 1에서, $J(\theta)$는 목적함수를 의미하며, π는 θ의 매개변수로 표현된 정책으로 θ의 변화에 따라 정책이 영향받는 것을 의미한다. A2C는 각 스텝이 종료될 때, 두 네트워크의 가중치를 업데이트하므로 s_t와 a_t는 t 스텝에서의 상태와 행동을 나타낸다. $A(s_t \mid a_t)$는 Advantage 함수로 시간차 오차 (temporal-difference error)이며

비평가 네트워크의 출력과 실제 보상의 차이 값이다. $\pi_\theta \log(a_t \mid s_t)$는 t 단계 상태에서의 행동을 결정하는 확률이다.

A2C는 정책 및 가치 기반 네트워크를 활용하여 학습의 안정성과 효율성을 증가시킨다. 경험 재현 메모리를 활용하는 DQN과 달리 각 에피소드의 경험을 즉시 사용하여 학습하므로 상대적으로 적은 메모리를 사용한다[7].

2. 셀 분해 알고리즘

셀 분해 알고리즘은 로봇 경로 계획 문제에 실제 환경의 특징을 이용하기 위해 사용되는 방법이다[8, 9]. 셀 분해는 연속적인 복잡한 환경을 더 작은 단위인 셀로 분할하여 공간 탐색 문제를 단순화시킬 수 있다.

그림 3은 근사 셀 분해(approximate cell decomposition)를 통해서 장애물이 존재하는 2차원 환경을 근사 셀로 분해한 모습을 보인다. 근사 셀 분해는 환경을 동일한 크기의 그리드로 분해한다. 분해된 그리드에 장애물이 포함되어 경로를 탐색하기 어려운 경우 더 작은 크기의 그리드로 분해한다.

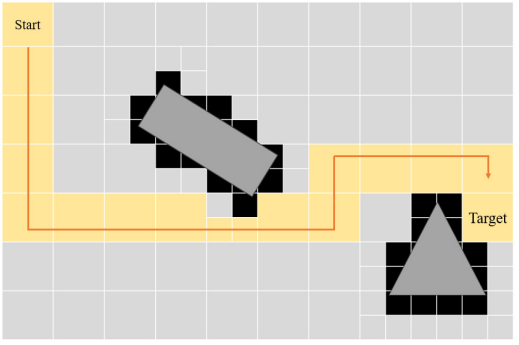

그림 3. 2차원 근사 셀 분해.
Fig. 3. 2-dimensional approximate cell decomposition.

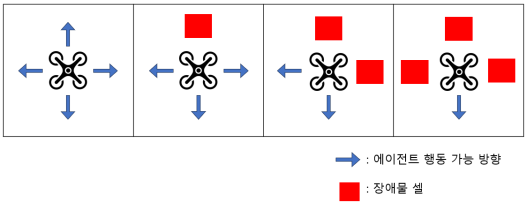

그림 4. 유효 행동 선택 전략.
Fig. 4. Effective action selection strategy.

Ⅲ. 시스템 모델

1. 환경 설계

학습 환경의 크기가 증가할수록 에이전트가 탐험해야 하는 상태 공간의 개수도 증가한다. 이는 에이전트가 특정 상태에서 가능한 행동의 수에 의해 학습의 복잡도가 기하급수적으로 증가하는 것을 의미한다. 에이전트는 최적의 정책을 얻기 위해 많은 상태 공간을 탐색하는 과정에서, 과도한 탐색으로 인해 성능이 저하되고 학습 수렴성을 기대하기 힘들다. 이를 해결하기 위해 학습 환경을 셀 분해 방법을 통해 학습 복잡도를 완화시켰다.

학습 환경을 근사 셀로 분해하여 안전한 셀과 장애물이 존재하는 셀로 구분한다. 에이전트가 현재 상태를 환경에 전달하면, 환경은 에이전트가 장애물 셀로 이동할 수 있는 행동을 제한한다. 그림 4에서 에이전트는 2차원(앞, 뒤, 오른쪽, 왼쪽)의 가능한 행동을 가진다. 환경은 현재 상태에서 유효한 행동들을 에이전트의 행동 선택 과정에 전달하여, 장애물이 존재하는 셀로 이동하는 행동을 선택지에서 제외할 수 있게 한다. 이를 통해 탐험해야 하는 환경의 규모를 축소하여 장애물 회피와 에이전트의 탐험 공간을 축소할 수 있다.

2. 보상 함수 설계

보상 함수 설계는 강화학습 에이전트가 최적 행동 학습에 중요한 역할을 한다. 고차원 환경에서 최단 경로를 위해 -1의 보상을 설정하는 것은 희소 보상 문제를 발생시켜 학습이 이루어지지 않을 수 있다. 학습의 수렴성과 행동 선택에 따른 선악을 에이전트가 반영할 수 있도록 본 논문에서는 상태에 따라 (-1, 1] 범위의 값을 가지도록 보상 함수를 설계하였다.

에이전트의 시작 위치(s_0)와 목표 위치(s_{target}) 사이의 맨해튼 거리를 $d_{\max}$로 설정하고, 에이전트의 현재 위치(s_t)와 목표 위치의 거리를 d_t, 다음 위치(s_{t+1})와 목표 위치의 거리를 d_{t+1}으로 설정한다. 수식 2, 수식 3, 그리고 수식 4과 같이 표현할 수 있다.

$$d_{\max} = \| s_0 - s_{target} \|, \quad (2)$$

$$d_t = \| s_t - s_{target} \|, \quad (3)$$

$$d_{t+1} = \| s_{t+1} - s_{target} \|. \quad (4)$$

쌍곡선 탄젠트를 이용하여 행동에 따른 보상을 (-1, 1) 범위로 설정한다. 이를 적용한 식은 수식 5와 같다.

$$R_a = \left|\tanh\left(d_{\max} - d_t\right)\right| \times \frac{d_t - d_{t+1}}{\left|d_t - d_{t+1}\right|}, \quad (5)$$

$$R = \begin{pmatrix} 1 & where\ s_t = target \\ R_a & other \end{pmatrix}. \quad (6)$$

수식 6은 전체 보상 함수이다. 에이전트가 현재 상태에서 다음 상태로 이동했을 때, d_{t+1}이 d_t에 비해서 감소하면 R_a는 양수의 보상을 받고, 그 반대의 경우 음수의 보상을 받는다. 보상 함수는 수식 6과 같이 설정했다. 목표 지점에 도달할 때는 1, 그 외의 행동은 R_a로 설정했다.

표 1. 시뮬레이션 파라미터.
Table 1. Simulation parameters.

구분	학습 파라미터
환경 크기	(100, 100, 100)
셀 분해 크기	10
행동	6 (상, 하, 좌, 우, 전, 후)
학습 에피소드	10,000
에피소드 당 최대 스텝	1,000
감가율 (Discount factor)	0.99
최적화 함수(Optimizer)	Adam
액터 네트워크 학습률	1e-6
비평자 네트워크 학습률	1e-5

Ⅳ. 시뮬레이션 결과

1. 시뮬레이션 파라미터 설정

학습 환경은 (100, 100, 100) 크기의 장애물 환경으로 설정했다. 셀 분해를 통해 장애물이 존재하는 셀과 안전한 셀로 구분하고, 유효한 행동 선택을 위해 이산-행동 공간으로 설계하였다. 시뮬레이션 파라미터는 표 1과 같다.

셀 분해를 위한 셀 분해 크기는 10으로 설정했다. 최대 학습 에피소드는 10,000번, 에피소드 당 최대 스텝(행동) 횟수는 1,000번으로 설정했다. 감가율은 0.99로 설정했다. 네트워크의 최적화 함수는 Adam을 사용하였고, 액터 네트워크의 학습률은 1e-6, 비평자 네트워크 학습률은 1e-5로 설정하였다.

본 논문에서 제안한 모델의 학습 효율성을 평가하기 위해서 제안한 모델과 동일한 파라미터를 갖는 A2C 알고리즘 비교하였다. 제안한 모델에서는 셀 분해 기법을 통해 얻은 유효한 행동이 에이전트에 제공되는 반면, 비교 모델에서는 유효한 행동이 에이전트에게 전달되지 않는다. 본 논문에서는 제안한 모델을 셀 분해 에이전트라 명명하고, 비교 모델을 셀 미분해 에이전트로 명명하여 두 모델의 성능을 비교한다.

2. 시뮬레이션 분석

Intel 13th i7-13700F CPU와 64GB RAM, RTX 4070 12GB VGA을 사용하여 셀 분해 에이전트는 10,000번의 에피소드 학습에 약 40분이 소요되었고, 셀 미분해 에이전트는 약 4시간이 소요되었다. 그림 5는 셀 분해 에이전트와 셀 미분해 에이전트의 에피소드에 에피소드 진행에 따른 보상의 변화를 보여준다. 셀 분해 에이전트는 약 100번째 에피소드에서 보상이 최대로 수렴한 것을 확인할 수 있다.

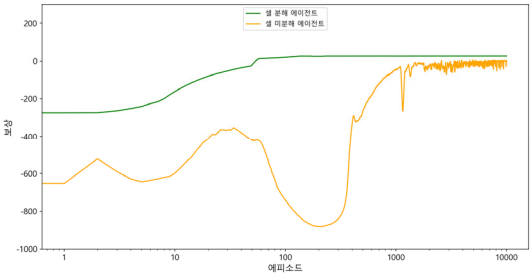

그림 5. 에이전트들의 보상 그래프.
Fig. 5. Reward graph of agents.

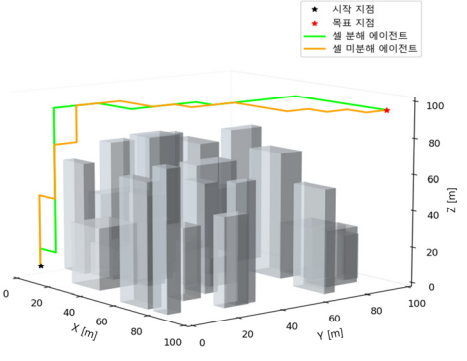

그림 6. 학습 환경에서의 에이전트 경로 계획 결과.
Fig. 6. Results in path planning in the learning environment.

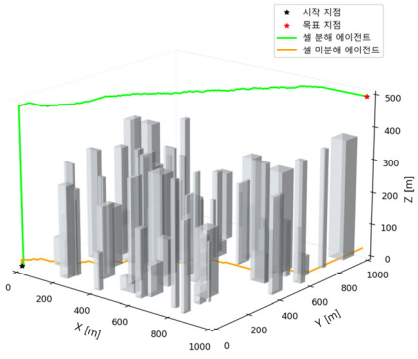

그림 7. 새로운 환경에서 에이전트 경로 계획 결과.
Fig. 7. Results of path planning in the new environment.

표 2. 각 환경에서 경로 계획에 움직인 횟수.
Table 2. Step counts in the path plan in each environment.

	셀 분해 에이전트	셀 미분해 에이전트
그림 6 환경	27회	50회
그림 7 환경	247회	1,000회

셀 미분해 에이전트의 경우 에피소드가 진행되면서 보상의 변화가 급격하게 변화하는 모습이 확인되었으며, 에피소드의 마지막 부분에서 분산되는 것을 확인하였다. 셀 분해 에이전트가 셀 미분해 에이전트에 비교하여 학습의 수렴성이 높고, 분산이 적다는 것을 확인할 수 있다.

그림 6, 7과 표 2는 두 에이전트가 각각의 환경에서 시작 지점과 목표 지점까지의 경로를 계획한 그림과 경로를 계획할 때 이동한 스텝 횟수를 나타낸다.

그림 6의 환경은 두 에이전트가 학습할 때 경험한 환경이다. 셀 분해 에이전트는 27회의 움직임, 셀 미분해 에이전트는 50회의 움직임으로 경로를 완성했다.

그림 7은 환경 크기를 (1000, 1000, 500)으로 증가하고, 장애물의 배치와 개수가 변화된 환경에서의 시뮬레이션 결과이다. 셀 분해 에이전트는 247회의 움직임으로 경로를 계획했지만, 셀 미분해 에이전트는 최대 스텝 1,000회에 도달하며 경로를 완성하지 못했다.

셀 분해 에이전트는 셀 미분해 에이전트보다 적은 움직임으로 경로를 계획하였고, 고차원 환경에서 유연성과 환경에 대한 적응성을 보여준다. 이를 통해 셀 분해가 적용된 환경의 강화학습 모델이 일반적인 환경의 강화학습 모델에 비해 높은 학습 속도를 가지며 환경에 대한 높은 유연성과 수렴성을 가지는 것을 확인할 수 있다.

V. 결 론

강화학습 기반 경로 계획 알고리즘은 고차원 환경에서 훈련 복잡도가 증가하여 수렴성을 기대하기 힘든 한계가 있다. 본 논문에서 제안된 모델은 강화학습에서 환경의 복잡도를 줄이기 위한 셀 분해 방법을 제시한다. 제안한 모델은 환경을 세부적으로 분해하여 에이전트에게 유효한 선택지를 제공하여, 학습 과정의 복잡도를 감소시키고 에이전트의 학습 효율성을 개선 시킬 수 있음을 확인하였다. 제안한 모델은 환경 크기의 변화, 장애물의 위치 변화 등 새로운 환경에서 비행경로를 생성할 수 있는 유효성을 입증하였다.

본 논문에서 제안한 모델을 활용하여 동적인 장애물이 포함된 복합 환경에서 환경의 복잡도를 낮추어 훈련 복잡도를 개선하고 강화학습 기반 경로 계획 알고리즘의 수렴성에 기여할 것으로 예상된다.

References

[1] S. Aggarwal and N. Kumar, "Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges," Computer Communications, vol. 149, pp. 270-299, Jan. 2020.
DOI: https://doi.org/10.1016/j.comcom.2019.10.014

[2] K.-T. Kim and G.-W. Jeon, "A Path Planning to Maximize Survivability for Unmanned Aerial Vehicle based on 3-dimensional Environment," IE interfaces, vol. 24, no. 4, pp. 304-313, Dec. 2011.
DOI: https://doi.org/10.7232/IEIF.2011.24.4.304

[3] H. Yang, H. Li, K. Liu, W. Yu, and X. Li, "Research on path planning based on improved RRT-Connect algorithm," in 2021 33rd Chinese Control and Decision Conference (CCDC), pp. 5707-5712, May, 2021.
DOI: https://doi.org/10.1109/CCDC52312.2021.9602203

[4] C. Chronis, G. Anagnostopoulos, E. Politi, A. Garyfallou, I. Varlamis, and G. Dimitrakopoulos, "Path planning of autonomous UAVs using reinforcement learning," J. Phys.: Conf. Ser., vol. 2526, no. 1, p. 012088, Jun. 2023.
DOI: https://doi.org/10.1088/1742-6596/2526/1/012088

[5] C. Qi, C. Wu, L. Lei, X. Li, and P. Cong, "UAV path planning based on the improved PPO algorithm," in 2022 Asia Conference on Advanced Robotics, Automation, and Control Engineering (ARACE), pp. 193-199, Aug. 2022.
DOI: https://doi.org/10.1109/ARACE56528.2022.00040

[6] K. H. Kim et al., "Research on Unmanned Aerial Vehicle Mobility Model based on Reinforcement Learning," The Journal of The Institute of Internet,

Broadcasting and Communication, vol. 23, no. 6, pp. 33-39, Dec. 2023.
DOI: https://doi.org/10.7236/JIIBC.2023.23.6.33

[7] Alibabaei, Khadijeh, Pedro D. Gaspar, Eduardo Assunção, Saeid Alirezazadeh, Tânia M. Lima, Vasco N. G. J. Soares, and João M. L. P. Caldeira. "Comparison of On-Policy Deep Reinforcement Learning A2C with Off-Policy DQN in Irrigation Optimization: A Case Study at a Site in Portugal." Computers, vol. 11, no. 7, July, 2022.
DOI: https://doi.org/10.3390/computers11070104.

[8] H. Zhang, W. Lin, and A. Chen, "Path Planning for the Mobile Robot: A Review," Symmetry, vol. 10, no. 10, Art. no. 10, Oct. 2018.
DOI: https://doi.org/10.3390/sym10100450

[9] F. Samaniego, J. Sanchis, S. García-Nieto, and R. Simarro, "UAV motion planning and obstacle avoidance based on adaptive 3D cell decomposition: Continuous space vs discrete space," in 2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM), pp. 1-6, Oct. 2017.
DOI: https://doi.org/10.1109/ETCM.2017.8247533

[10] J. Wang, Z. Zhao, J. Qu, and X. Chen, "APPA-3D: an autonomous 3D path planning algorithm for UAVs in unknown complex environments," Sci Rep, vol. 14, no. 1, Art. no. 1, Jan. 2024.
DOI: https://doi.org/10.1038/s41598-024-51286-2

저 자 소 개

김 경 훈(준회원)

- 2024년 2월 : 광운대학교 전자융합공학 학사 졸업
- 2024년 3월 ~ 현재 : 광운대학교 전자융합공학과 석박사통합과정
- 관심분야 : 딥러닝, 강화학습, 자율주행, 측위 시스템

황 병 선(준회원)

- 2023년 2월 : 광운대학교 전자융합공학 학사 졸업
- 2023년 3월 ~ 현재 : 광운대학교 전자융합공학과 석박사통합과정
- 관심분야 : 측위 시스템, 인공지능, 무선 통신 시스템

선 준 호(준회원)

- 2021년 2월 : 광운대학교 전자융합공학 학사 졸업
- 2021년 3월 ~ 현재 : 광운대학교 전자융합공학과 석박사통합과정
- 관심분야 : 딥러닝, 이상 탐지, 스마트 그리드, 위성통신

김 수 현(준회원)

- 2019년 2월 : 광운대학교 전자융합공학과 졸업
- 2019년 3월 ~ 현재 : 광운대학교 전자융합공학과 석박통합과정
- 관심분야 : 차세대이동통신, 인공지능, 스마트 그리드

김 진 영(정회원)

- 1998년 2월 : 서울대학교 전자공학과 공학박사
- 2001년 2월 : SK텔레콤 네트워 연구소 책임연구원
- 2001년 3월 ~ 현재 : 광운대학교 전자융합공학과 교수
- 관심분야 : 인공지능, 차세대이동통신, 전력선통신, 가시광통신, 무선 측위 시스템

※ 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2021-0-00892-004, 우주공간의 편재 지능을 위한 저궤도 위성통신 시스템의 핵심 물리 계층 기술 연구)과 2024년도 광운대학교 우수연구자 지원 사업에 의해 연구되었음