HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Nguyễn Thanh Hùng

NGHIÊN CỨU MÔ HÌNH HỌC TĂNG CƯỜNG ĐA TÁC TỬ TRONG BÀI TOÁN ĐIỀU KHIỂN TÍN HIỆU GIAO THÔNG TẠI VIỆT NAM

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2026

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Nguyễn Thanh Hùng

NGHIÊN CỨU MÔ HÌNH HỌC TĂNG CƯỜNG ĐA TÁC TỬ TRONG BÀI TOÁN ĐIỀU KHIỂN TÍN HIỆU GIAO THÔNG TẠI VIỆT NAM

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN
MÃ SỐ: 8.48.01.04 (Hệ thống thông tin)

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: **TS. NGUYỄN ĐÌNH HOÁ**

HÀ NỘI – 2026

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong đề án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Các tài liệu tham khảo trích dẫn trong đề án đều được ghi rõ nguồn gốc. Mọi sự giúp đỡ cho việc thực hiện đề án này đã được cảm ơn và ghi nhận.

Hà Nội, ngày 28 tháng 7 năm 2026

Học viên thực hiện

(Ký và ghi rõ họ tên)

Hùng

Nguyễn Thanh Hùng

Lời đầu tiên, em xin bày tỏ lòng biết ơn sâu sắc tới **TS. Nguyễn Đình Hoá**, người đã tận tình hướng dẫn, chỉ bảo và định hướng cho em trong suốt quá trình thực hiện đề án này. Những kiến thức và kinh nghiệm quý báu mà Thầy truyền đạt là hành trang quan trọng giúp em hoàn thành tốt đề tài.

Em cũng xin gửi lời cảm ơn trân trọng đến Khoa Đào tạo Sau đại học cùng toàn thể các thầy cô giáo tại **Học viện Công nghệ Bưu chính Viễn thông**. Môi trường học tập chuyên nghiệp và sự tận tâm của các thầy cô đã trang bị cho em nền tảng kiến thức vững chắc để thực hiện nghiên cứu này.

Cuối cùng, em xin gửi lời cảm ơn đến gia đình, bạn bè và các đồng nghiệp đã luôn ở bên động viên, khích lệ và tạo mọi điều kiện thuận lợi nhất để em có thể yên tâm học tập và hoàn thành chương trình đào tạo.

Mặc dù đã rất cố gắng, song đề án khó tránh khỏi những thiếu sót. Em rất mong nhận được sự đóng góp ý kiến của quý Thầy, Cô và các bạn để đề tài được hoàn thiện hơn.

Em xin chân thành cảm ơn!

MỤC LỤC

LỜI CAM ĐOAN

LỜI CẢM ƠN

MỤC LỤC i

DANH SÁCH TỪ VIẾT TẮT iv

DANH SÁCH BẢNG vii

DANH SÁCH HÌNH VẼ viii

MỞ ĐẦU 1

Chương 1 TỔNG QUAN VỀ HỌC TĂNG CƯỜNG TRONG ĐIỀU KHIỂN TÍN HIỆU GIAO THÔNG 4

1.1 Các phương pháp dựa trên mô hình toán học 4

1.1.1 Điều khiển chu kỳ cố định 5

1.1.2 Điều khiển theo nhu cầu tại nút giao đơn lẻ 6

1.1.3 Điều khiển áp lực cực đại 6

1.1.4 Điều khiển thích ứng quy mô mạng 8

1.2 Các phương pháp học từ dữ liệu 8

1.2.1 Phân tích và dự báo sử dụng học máy truyền thống 8

1.2.2 Nhận biết trạng thái giao thông với học sâu 10

1.2.3 Điều khiển trực tiếp với học tăng cường 10

1.2.4 Học tăng cường và PPO 11

1.2.5 Sự phát triển của cơ chế phối hợp và quy mô mạng lưới 12

1.3 Học tăng cường đa tác tử trong giao thông 12

1.3.1 Khung lý thuyết Markov đa tác tử 12

1.3.2 Thách thức trong điều khiển đa tác tử 13

1.3.3 Chiến lược phối hợp 13

1.4 Mạng nơ-ron đồ thị cho bài toán phối hợp 14

1.4.1 Biểu diễn đồ thị giao thông 14

1.4.2 Cơ chế truyền tin và sự chú ý 15
1.4.3 Khả năng mở rộng và điều khiển mạng lưới lớn 15
1.5 Phân tích khoảng trống nghiên cứu 15
1.5.1 Sự thích nghi với luồng giao thông hỗn hợp 16
1.5.2 Hạ tầng dữ liệu và môi trường mô phỏng 16
1.5.3 Tính phối hợp giữa các nút giao trong mạng lưới dày đặc . . . 17
1.6 Kết luận chương 17

Chương 2 MÔ HÌNH HỌC TĂNG CƯỜNG ĐA TÁC TỬ TRONG ĐIỀU KHIỂN TÍN HIỆU GIAO THÔNG 19

2.1 Nền tảng học tăng cường đa tác tử 19
2.1.1 Từ MDP đến trò chơi Markov 19
2.1.2 Huấn luyện tập trung, thực thi phân tán 19
2.1.3 Các thách thức chính trong MARL 20
2.2 Cơ chế cập nhật chính sách điều khiển 20
2.2.1 Giới hạn mức thay đổi bằng PPO 20
2.2.2 Ước lượng lợi thế bằng GAE 21
2.2.3 Phương pháp sử dụng PPO và GAE trong quá trình huấn luyện 22
2.3 Mô hình hóa bài toán 22
2.3.1 Trò chơi Markov 22
2.3.2 Không gian quan sát 23
2.3.3 Không gian hành động và cơ chế lọc hành động 25
2.3.4 Ràng buộc chuyển pha an toàn 26
2.3.5 Hàm phần thưởng 26
2.4 Môi trường mô phỏng 27
2.4.1 Mạng lưới và kịch bản 27
2.4.2 Sinh nhu cầu giao thông từ dữ liệu UTM 28
2.4.3 Hệ số PCU và loại phương tiện 28
2.4.4 Tác tử điều khiển và bộ điều khiển cố định 29
2.5 Đồ thị chính sách 30
2.5.1 Xây dựng đồ thị 30
2.5.2 Cơ chế chú ý GAT 31
2.6 Kiến trúc các thuật toán 32
2.6.1 MAPPO 33

2.6.2 HAPPO 34
2.6.3 GAT-MAPPO 35
2.6.4 GAT-HAPPO 37
2.7 Thiết lập thông số huấn luyện 39
2.8 Giao thức lưu mô hình và bảo đảm tính tái lập 40
2.9 Minh họa cấu hình chương trình đèn tín hiệu trong SUMO 41
2.10 Kết luận chương 41

Chương 3 THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ **42**
3.1 Kịch bản và giao thức đánh giá 42
3.1.1 Kịch bản Thanh Xuân UTM 42
3.1.2 Các phương pháp so sánh 43
3.1.3 Giao thức huấn luyện và đánh giá 44
3.2 Chỉ số đánh giá 45
3.3 Kết quả đánh giá đa hạt giống 45
3.3.1 Bảng kết quả chính 45
3.3.2 So sánh với Max-Pressure 47
3.3.3 Vai trò của GAT 48
3.4 Diễn giải theo từng chỉ số 50
3.4.1 Thời gian chờ và tốc độ trung bình 50
3.4.2 Số phương tiện hoàn thành hành trình 50
3.4.3 Dịch chuyển bắt buộc 50
3.5 Nhận xét kết quả 50
3.6 Hạn chế 51
3.7 Kết luận chương 52

KẾT LUẬN **53**

TÀI LIỆU THAM KHẢO **56**

DANH SÁCH TỪ VIẾT TẮT

TT	Từ viết tắt	Tiếng Anh	Tiếng Việt
1	**AI**	Artificial Intelligence	Trí tuệ nhân tạo
2	**API**	Application Programming Interface	Giao diện lập trình ứng dụng
3	**CNN**	Convolutional Neural Network	Mạng nơ-ron tích chập
4	**CoLight**	CoLight	Mô hình CoLight
5	**CPU**	Central Processing Unit	Bộ xử lý trung tâm
6	**CTDE**	Centralized Training - Decentralized Execution	Huấn luyện tập trung - thực thi phân tán
7	**DQN**	Deep Q-Network	Mạng Q sâu
8	**DRL**	Deep Reinforcement Learning	Học tăng cường sâu
9	**FRAP**	Fair Resource Allocation for Phases	Phương pháp tối ưu pha đèn
10	**GAE**	Generalized Advantage Estimation	Ước lượng lợi thế tổng quát
11	**GAT**	Graph Attention Network	Mạng chú ý đồ thị
12	**GAT-MAPPO**	Graph Attention Multi-Agent PPO	PPO đa tác tử kết hợp cơ chế chú ý đồ thị
13	**GCN**	Graph Convolutional Network	Mạng tích chập đồ thị
14	**GNN**	Graph Neural Network	Mạng nơ-ron đồ thị
15	**GPU**	Graphical Processing Unit	Bộ xử lý đồ họa
16	**HAPPO**	Heterogeneous-Agent PPO	Tối ưu hóa chính sách tác tử không đồng nhất
17	**IntelliLight**	IntelliLight	Mô hình IntelliLight
18	**IQL**	Independent Q-Learning	Học Q độc lập
19	**ITS**	Intelligent Transportation Systems	Hệ thống giao thông thông minh
20	**KNN**	K-Nearest Neighbors	Thuật toán K lân cận gần nhất

21	**LSTM**	Long Short-Term Memory	Mạng bộ nhớ dài-ngắn hạn
22	**MADDPG**	Multi-Agent Deep Deterministic Policy Gradient	Thuật toán điều khiển đa tác tử
23	**MAPPO**	Multi-Agent PPO	PPO đa tác tử
24	**MARL**	Multi-Agent Reinforcement Learning	Học tăng cường đa tác tử
25	**MAT**	Multi-Agent Transformer	Transformer đa tác tử
26	**MDP**	Markov Decision Process	Quá trình quyết định Markov
27	**MLP**	Multi-Layer Perceptron	Perceptron đa lớp
28	**MPLight**	MPLight	Mô hình MPLight
29	**OSM**	OpenStreetMap	Bản đồ mở OpenStreetMap
30	**PCU**	Passenger Car Unit	Đơn vị xe con quy đổi
31	**PPO**	Proximal Policy Optimization	Tối ưu hóa chính sách cận gần
32	**PressLight**	PressLight	Mô hình PressLight
33	**QMIX**	Q-Mixer Value Decomposition	Phân rã giá trị QMIX
34	**RL**	Reinforcement Learning	Học tăng cường
35	**RNN**	Recurrent Neural Network	Mạng nơ-ron tái hồi
36	**SCATS**	Sydney Coordinated Adaptive Traffic System	Hệ thống điều khiển giao thông thích ứng Sydney
37	**SCOOT**	Split Cycle Offset Optimisation Technique	Hệ thống tối ưu chu kỳ và độ lệch
38	**SOTA**	State-of-the-Art	Tiên tiến nhất
39	**STGCN**	Spatio-Temporal Graph Convolutional Network	Mạng tích chập đồ thị không-thời gian
40	**SUMO**	Simulation of Urban MObility	Mô phỏng di động đô thị
41	**SVM**	Support Vector Machine	Máy vectơ hỗ trợ
42	**SVR**	Support Vector Regression	Hồi quy vectơ hỗ trợ
43	**TraCI**	Traffic Control Interface	Giao diện điều khiển giao thông

44	**UTM**	Urban Traffic Management	Quản lý giao thông đô thị

DANH SÁCH BẢNG

Bảng 1.1 So sánh các mô hình học tăng cường sâu trong điều khiển đèn tín hiệu 11

Bảng 2.1 So sánh kiến trúc bốn biến thể thuật toán được triển khai 39
Bảng 2.2 Các thông số chính trong cấu hình huấn luyện 40

Bảng 3.1 Cấu hình kịch bản thực nghiệm Thanh Xuân UTM 42
Bảng 3.2 Kết quả trung bình ± độ lệch chuẩn trên 10 hạt giống đánh giá, kịch bản Thanh Xuân UTM 46
Bảng 3.3 Độ chênh lệch trung bình so với Max-Pressure 47

DANH SÁCH HÌNH VẼ

Hình 1.1 Sơ đồ chu kỳ đèn tín hiệu 5
Hình 1.2 Cơ chế tính áp lực dựa trên chênh lệch hàng đợi giữa làn tiếp cận và làn thoát 7
Hình 1.3 Minh họa ánh xạ mạng giao thông thành đồ thị điều khiển 14
Hình 1.4 So sánh dòng xe quy chuẩn theo làn và dòng xe hỗn hợp thực tế tại Việt Nam 16

Hình 2.1 Sơ đồ tương tác giữa tác tử và môi trường trong mô hình trò chơi Markov 23
Hình 2.2 Quy trình trích xuất và mã hóa vector quan sát cục bộ dựa trên hệ số PCU 25
Hình 2.3 Lưu đồ chuẩn bị dữ liệu đầu vào cho kịch bản thực nghiệm . . . 28
Hình 2.4 Minh họa một bộ điều khiển đèn tín hiệu và các pha thuộc cùng một tác tử điều khiển 30
Hình 2.5 Cơ chế truyền tin trong GAT với trọng số chú ý động giữa các nút giao trong đồ thị chính sách 31
Hình 2.6 So sánh cơ chế actor và cập nhật của MAPPO và HAPPO 33
Hình 2.7 So sánh kiến trúc và cơ chế truyền tin đồ thị trong GAT-MAPPO và GAT-HAPPO 36

Hình 3.1 Mạng lưới giao thông khu vực Thanh Xuân 43
Hình 3.2 So sánh bốn chỉ số đánh giá chính trên kịch bản Thanh Xuân UTM 47
Hình 3.3 Thời gian chờ trung bình theo từng hạt giống đánh giá 49
Hình 3.4 Số phương tiện hoàn thành hành trình theo từng hạt giống đánh giá trên kịch bản Thanh Xuân UTM 49

MỞ ĐẦU

Tốc độ đô thị hóa nhanh cùng với sự gia tăng nhanh chóng của phương tiện cá nhân đang đặt ra nhiều thách thức cho hệ thống giao thông tại các đô thị lớn ở Việt Nam, đặc biệt là Hà Nội và Thành phố Hồ Chí Minh. Ùn tắc giao thông không chỉ gây thiệt hại về kinh tế do lãng phí thời gian và nhiên liệu, mà còn ảnh hưởng tiêu cực đến chất lượng sống của người dân đô thị.

Các hệ thống điều khiển đèn tín hiệu hiện nay ở Việt Nam vẫn chủ yếu dựa trên các chu kỳ cố định được thiết lập từ kết quả khảo sát lưu lượng định kỳ. Trong khi đó, lưu lượng giao thông thực tế thường biến động theo thời gian trong ngày, điều kiện thời tiết và các sự kiện đột xuất. Do không phản ứng trực tiếp với những biến động này, điều khiển cố định có thể duy trì đèn xanh ở hướng có ít phương tiện, trong khi các hướng khác đang xuất hiện tình trạng ùn tắc.

Trong bối cảnh đó, học tăng cường đa tác tử (*Multi-Agent Reinforcement Learning* - MARL) là một hướng tiếp cận có tiềm năng đối với bài toán điều khiển tín hiệu giao thông. Theo cách tiếp cận này, mỗi bộ điều khiển đèn tín hiệu có nhiều hơn một giai đoạn xanh hợp lệ được xem là một tác tử điều khiển. Các tác tử học chính sách điều khiển thông qua tương tác với môi trường mô phỏng và có thể phối hợp với các tác tử lân cận để cải thiện trạng thái vận hành của mạng lưới. Tuy nhiên, việc áp dụng trực tiếp các mô hình MARL đã được nghiên cứu trong bối cảnh quốc tế vào Việt Nam vẫn gặp nhiều rào cản. Những rào cản này xuất phát từ đặc thù dòng xe hỗn hợp với tỷ lệ xe máy cao, hành vi di chuyển không tuân thủ làn đường nghiêm ngặt và sự thiếu hụt dữ liệu nhu cầu giao thông được chuẩn hóa cho từng khu vực đô thị.

Từ thực tế trên, đề tài **"Nghiên cứu mô hình học tăng cường đa tác tử trong bài toán điều khiển tín hiệu giao thông tại Việt Nam"** được thực hiện nhằm thu hẹp khoảng cách giữa các mô hình MARL hiện đại và điều kiện giao thông đặc thù tại Việt Nam. Đề án tập trung xây dựng quy trình mô phỏng vi mô từ dữ liệu khảo sát giao thông thực tế, đánh giá định lượng bốn thuật toán MARL có khả năng triển khai và so sánh các thuật toán này với các phương pháp đối chứng trên một mạng lưới giao thông thực tế tại Hà Nội.

Phạm vi của đề án được giới hạn trong môi trường mô phỏng vi mô trên nền tảng mô phỏng di động đô thị (*Simulation of Urban MObility* - SUMO) [1] cho khu vực Thanh Xuân, Hà Nội. Nhu cầu giao thông được sinh từ dữ liệu khảo sát UTM và

được điều chỉnh theo kịch bản mục tiêu. Đề án không khẳng định khả năng triển khai ngoài thực địa và cũng chưa đánh giá trên dữ liệu camera thời gian thực. Các kết quả trong nghiên cứu này vì vậy được diễn giải trong phạm vi mô phỏng và giao thức thực nghiệm đã xác định.

Mục tiêu của đề án là nghiên cứu khả năng áp dụng học tăng cường đa tác tử trong bài toán điều khiển tín hiệu giao thông trên mạng lưới đô thị có dòng xe hỗn hợp. Để thực hiện mục tiêu này, đề án tập trung xây dựng một môi trường mô phỏng giao thông cho khu vực Thanh Xuân, Hà Nội trên nền tảng SUMO, trong đó mạng đường được lấy từ dữ liệu bản đồ mở OpenStreetMap (*OpenStreetMap* - OSM) và nhu cầu giao thông được xây dựng từ dữ liệu khảo sát UTM-Hanoi. Trên môi trường này, đề án đặc tả bài toán điều khiển dưới dạng hệ đa tác tử, thiết kế không gian quan sát, không gian hành động, hàm phần thưởng và cơ chế phối hợp giữa các nút giao. Các thuật toán MARL được triển khai và đánh giá định lượng dựa trên các chỉ số như số phương tiện hoàn thành hành trình, thời gian chờ trung bình, tốc độ trung bình và số sự kiện dịch chuyển bắt buộc.

Từ các mục tiêu trên, đóng góp chính của đề án nằm ở khía cạnh ứng dụng và thực nghiệm. Trước hết, đề án xây dựng được một quy trình mô phỏng vi mô cho khu vực Thanh Xuân dựa trên dữ liệu bản đồ mở và dữ liệu khảo sát thực tế, có xét đến đặc trưng giao thông hỗn hợp thông qua hệ số đơn vị xe con quy đổi (*Passenger Car Unit* - PCU) và mô hình chuyển làn phụ trong SUMO. Bên cạnh đó, đề án đặc tả một môi trường điều khiển đèn tín hiệu đa tác tử gồm tập tác tử điều khiển, vector quan sát cục bộ, không gian hành động có ràng buộc, cơ chế lọc hành động và hàm phần thưởng phản ánh các yếu tố bất lợi của trạng thái giao thông. Trên cơ sở môi trường này, đề án triển khai và so sánh bốn biến thể học tăng cường đa tác tử dựa trên PPO, gồm MAPPO, HAPPO, GAT-MAPPO và GAT-HAPPO, trong cùng một giao thức thực nghiệm. Kết quả đánh giá đa hạt giống với các phương pháp đối chứng như điều khiển cố định và Max-Pressure cung cấp bằng chứng thực nghiệm về hiệu quả và giới hạn của từng phương pháp trong kịch bản Thanh Xuân UTM, thay vì giả định một thuật toán vượt trội trên mọi tiêu chí.

Nội dung của đề án được trình bày trong ba chương chính. Chương 1 trình bày tổng quan về học tăng cường trong điều khiển tín hiệu giao thông. Chương này giới thiệu bối cảnh nghiên cứu, phân tích các phương pháp điều khiển truyền thống và hiện đại, đồng thời trình bày các nền tảng liên quan đến học tăng cường (*Reinforcement Learning* - RL), MARL và mạng nơ-ron đồ thị (*Graph Neural Network* - GNN). Trên

cơ sở đó, chương làm rõ khoảng trống nghiên cứu đối với bài toán điều khiển tín hiệu giao thông trong bối cảnh giao thông hỗn hợp và dữ liệu khảo sát tại Việt Nam.

Chương 2 trình bày mô hình học tăng cường đa tác tử cho điều khiển tín hiệu giao thông. Chương này mô tả quy trình xây dựng kịch bản mô phỏng từ dữ liệu UTM, cách đặc tả không gian quan sát, không gian hành động và hàm phần thưởng của môi trường, cũng như kiến trúc của bốn thuật toán được triển khai. Ngoài ra, chương còn trình bày đồ thị chính sách, giao thức lưu mô hình và cơ chế ghi nhận cấu hình nhằm hỗ trợ việc tái lập kết quả.

Chương 3 trình bày thực nghiệm và kết quả đánh giá. Chương này mô tả kịch bản thực nghiệm chính trên mạng giao thông Thanh Xuân, hệ thống chỉ số đánh giá và kết quả so sánh sáu phương pháp trên nhiều hạt giống ngẫu nhiên. Từ các kết quả định lượng, chương phân tích hiệu quả của từng phương pháp theo từng chỉ số vận hành và thảo luận các hạn chế của nghiên cứu.

Chương 1
TỔNG QUAN VỀ HỌC TĂNG CƯỜNG TRONG ĐIỀU KHIỂN TÍN HIỆU GIAO THÔNG

Chương này trình bày tổng quan các hướng nghiên cứu liên quan đến điều khiển tín hiệu giao thông, từ các phương pháp truyền thống dựa trên mô hình toán học và quy tắc điều khiển đến các tiếp cận hiện đại sử dụng trí tuệ nhân tạo. Trọng tâm của phần tổng quan là làm rõ vai trò của học tăng cường, học tăng cường đa tác tử và mạng nơ-ron đồ thị trong bài toán điều khiển mạng lưới giao thông đô thị có tính động cao. Trên cơ sở đó, đề án phân tích khoảng trống nghiên cứu đối với bối cảnh giao thông hỗn hợp tại Việt Nam, nơi tỷ lệ xe máy cao và hành vi di chuyển không hoàn toàn tuân theo cấu trúc làn đường cố định.

Bài toán điều khiển tín hiệu giao thông có thể được hiểu là quá trình lựa chọn thời điểm chuyển đổi giữa các giai đoạn xanh tại mỗi nút giao. Tại một nút giao, các hướng di chuyển xung đột được nhóm thành những giai đoạn xanh khác nhau; bộ điều khiển đèn tín hiệu có nhiệm vụ quyết định duy trì giai đoạn hiện tại hoặc chuyển sang giai đoạn tiếp theo nhằm cải thiện luồng phương tiện đi qua nút giao. Một chu kỳ đèn bao gồm thứ tự các giai đoạn xanh cùng với các khoảng thời gian vàng và đỏ toàn bộ giữa chúng. Khi bài toán được mở rộng từ một nút giao đơn lẻ sang mạng lưới nhiều nút giao liên kết chặt chẽ, quyết định tại một nút có thể ảnh hưởng trực tiếp đến dòng xe đến của các nút lân cận và làm phát sinh hiện tượng lan truyền ùn tắc. Vì vậy, các phương pháp điều khiển tín hiệu giao thông thường được xem xét theo hai nhóm chính: nhóm phương pháp truyền thống dựa trên mô hình toán học hoặc quy tắc điều khiển, và nhóm phương pháp học từ dữ liệu, bao gồm học máy, học sâu và học tăng cường.

1.1 Các phương pháp dựa trên mô hình toán học

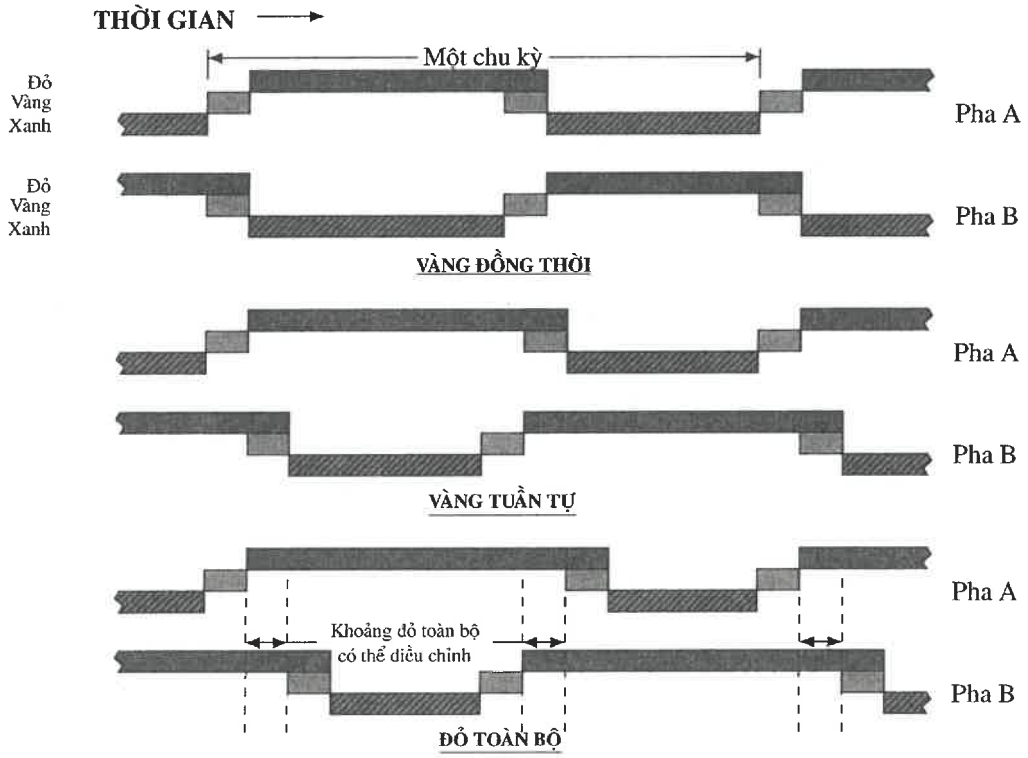
Các phương pháp dựa trên mô hình toán học là nhóm tiếp cận cổ điển trong điều khiển tín hiệu giao thông. Nhóm phương pháp này xây dựng quy tắc điều khiển dựa trên lý thuyết luồng giao thông, mô hình hàng đợi và các giả định về sự phân bố phương tiện trong mạng lưới. Trong phạm vi tổng quan của đề án, các phương pháp dựa trên mô hình toán học được trình bày thông qua bốn hướng tiêu biểu: điều khiển chu kỳ cố định, điều khiển theo nhu cầu tại nút giao đơn lẻ, điều khiển áp lực cực đại

và điều khiển thích ứng quy mô mạng.

1.1.1 Điều khiển chu kỳ cố định

Điều khiển chu kỳ cố định là một trong những phương pháp được sử dụng phổ biến nhất trong thực tế. Với phương pháp này, các tham số như độ dài chu kỳ, thứ tự pha và thời gian xanh được thiết lập trước dựa trên dữ liệu lịch sử hoặc kết quả khảo sát lưu lượng. Công thức của Webster [2, tr. 37] là một nền tảng quan trọng trong nhóm phương pháp này, vì công thức cho phép tính toán độ dài chu kỳ tối ưu (c_0) nhằm giảm thiểu tổng thời gian chậm trễ tại nút giao:

$$c_0 = \frac{1.5L + 5}{1 - Y} \tag{1.1}$$



Hình 1.1: Sơ đồ chu kỳ đèn tín hiệu (dựa trên mô hình Webster, [2])

Trong đó:

c_0 : Chu kỳ tối ưu, là thời gian chu kỳ mang lại độ trễ trung bình nhỏ nhất cho tất cả các phương tiện qua nút giao.

L : Tổng thời gian mất mát mỗi chu kỳ, là tổng thời gian mất mát của từng pha cộng với các khoảng thời gian mà tất cả các đèn đều đỏ hoặc đỏ vàng.

Y : Tổng tỷ số lưu lượng của toàn bộ nút giao, là tổng các giá trị y của các pha, trong đó $y = q/s$.

q : Lưu lượng luồng phương tiện, là số lượng xe trung bình đi qua một điểm theo một hướng trong một đơn vị thời gian.

s : Lưu lượng bão hòa, là tốc độ xả luồng tối đa của hàng đợi trong thời gian đèn xanh.

Mặc dù điều khiển chu kỳ cố định có ưu điểm là đơn giản và dễ triển khai, phương pháp này thiếu khả năng thích ứng với các biến động thời gian thực của lưu lượng giao thông. Hạn chế này trở nên rõ rệt trong các kịch bản giờ cao điểm, sự kiện bất thường hoặc sự cố đột xuất, khi lưu lượng thực tế có thể khác đáng kể so với dữ liệu lịch sử được dùng để thiết lập chu kỳ.

1.1.2 Điều khiển theo nhu cầu tại nút giao đơn lẻ

Điều khiển theo nhu cầu tại nút giao đơn lẻ sử dụng dữ liệu từ các cảm biến như vòng từ hoặc camera để phát hiện sự hiện diện của phương tiện trên từng làn đường. Dựa trên thông tin cảm biến, bộ điều khiển có thể điều chỉnh thời gian xanh theo nhu cầu thực tế tại nút giao. Ví dụ, nếu một hướng không có phương tiện chờ, pha đèn tương ứng có thể được kết thúc sớm để chuyển quyền ưu tiên sang hướng khác có nhu cầu cao hơn.

Phương pháp này có thể cải thiện hiệu năng cục bộ tại từng nút giao so với điều khiển chu kỳ cố định. Tuy nhiên, do quyết định chủ yếu dựa trên trạng thái tại một nút riêng lẻ, điều khiển theo nhu cầu vẫn gặp khó khăn khi cần phối hợp nhiều nút giao trên quy mô mạng lưới. Đặc biệt, việc duy trì các chiến lược phối hợp như "làn sóng xanh" trở nên phức tạp hơn khi độ dài chu kỳ và thời điểm chuyển pha thay đổi theo từng chu kỳ vận hành.

1.1.3 Điều khiển áp lực cực đại

Điều khiển áp lực cực đại là phương pháp thích ứng dựa trên lý thuyết mạng hàng đợi. Thay vì sử dụng chu kỳ cố định, phương pháp này lựa chọn trạng thái tín hiệu có áp lực lớn nhất để kích hoạt tại mỗi bước quyết định. Theo Varaiya [3, tr. 184], trọng số $w(l,m)$ của mỗi luồng di chuyển và áp lực $\gamma(S)$ của một ma trận điều khiển S được

tính như sau:

$$w(l,m) = x(l,m) - \sum_{p \in Out_m} r(m,p)x(m,p) \tag{1.2}$$

$$\gamma(S) = \sum_{l,m} c(l,m)w(l,m)S(l,m) \tag{1.3}$$

Trong đó:

$w(l,m)$: Trọng số của luồng di chuyển từ làn tiếp cận l sang làn thoát m.

$\gamma(S)$: Áp lực của ma trận điều khiển S.

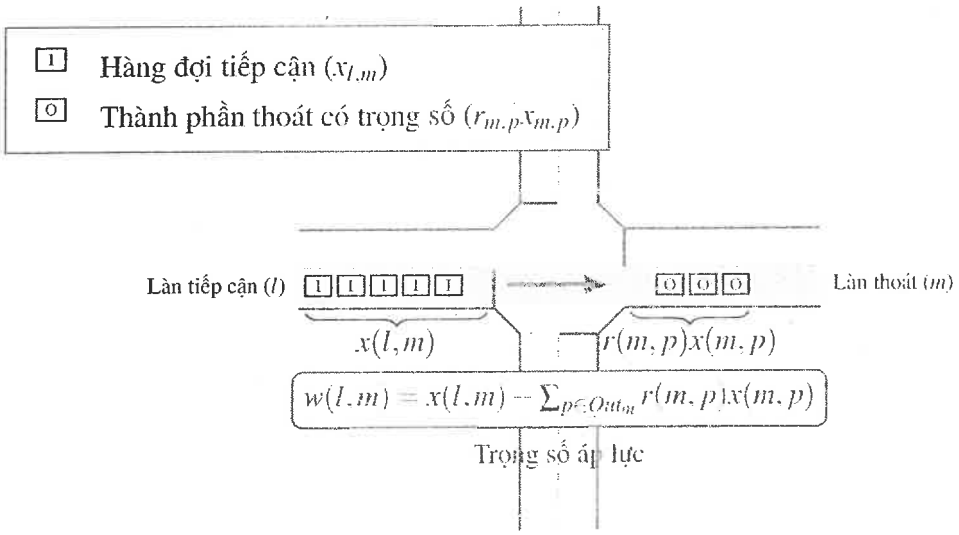
$x(l,m)$: Độ dài hàng đợi trên làn tiếp cận l cho luồng đi từ l sang m.

$x(m,p)$: Độ dài hàng đợi trên làn thoát m cho luồng tiếp tục sang làn kế tiếp p.

$r(m,p)$: Tỷ lệ chuyển hướng từ làn thoát m sang làn kế tiếp p.

$c(l,m)$: Lưu lượng bão hòa của luồng di chuyển từ làn tiếp cận l sang làn thoát m.

$S(l,m)$: Biến nhị phân, bằng 1 nếu luồng từ làn tiếp cận l sang làn thoát m được kích hoạt.



Hình 1.2: Cơ chế tính áp lực dựa trên chênh lệch hàng đợi giữa làn tiếp cận và làn thoát

Varaiya đã chứng minh rằng việc chọn S để tối đa hóa $\gamma(S)$ có thể tối đa hóa thông lượng của toàn mạng mà không cần biết chính xác tham số nhu cầu giao thông. Gần

đây, các nghiên cứu như PressLight [4] đã kế thừa nguyên lý áp lực cực đại và kết hợp nguyên lý này với học tăng cường sâu nhằm cải thiện hiệu suất điều khiển trong môi trường giao thông động.

1.1.4 Điều khiển thích ứng quy mô mạng

Các hệ thống như hệ thống tối ưu chu kỳ và độ lệch (*Split Cycle Offset Optimisation Technique* - SCOOT) và hệ thống điều khiển giao thông thích ứng Sydney (*Sydney Coordinated Adaptive Traffic System* - SCATS) đại diện cho nhóm điều khiển thích ứng quy mô mạng. Những hệ thống này thu thập dữ liệu từ mạng lưới cảm biến dày đặc và sử dụng các thuật toán tối ưu hóa trung tâm để điều chỉnh tham số điều khiển trên toàn mạng theo thời gian thực. Nhờ khả năng cập nhật tham số dựa trên trạng thái vận hành, điều khiển thích ứng quy mô mạng có thể phản ứng linh hoạt hơn so với điều khiển chu kỳ cố định.

Tuy nhiên, các hệ thống điều khiển thích ứng quy mô mạng thường đòi hỏi chi phí đầu tư và bảo trì hạ tầng cảm biến rất lớn. Bên cạnh đó, các mô hình toán học bên dưới các hệ thống này có thể gặp khó khăn khi mô tả dòng giao thông hỗn hợp không tuân theo các quy luật luồng xe truyền thống. Hạn chế này đặc biệt đáng chú ý trong bối cảnh các nước đang phát triển như Việt Nam, nơi xe máy chiếm tỷ lệ cao và hành vi di chuyển theo làn không luôn rõ ràng.

1.2 Các phương pháp học từ dữ liệu

Khác với nhóm phương pháp ở Mục 1.1, vốn chủ yếu dựa trên mô hình luồng giao thông, quy tắc điều khiển được thiết kế trước hoặc dữ liệu cảm biến tức thời, các phương pháp học từ dữ liệu tập trung vào việc học quan hệ giữa trạng thái giao thông và quyết định điều khiển từ dữ liệu quan sát được. Sự phát triển của nhóm phương pháp này gắn liền với quá trình ứng dụng trí tuệ nhân tạo vào giao thông, bắt đầu từ các kỹ thuật học máy truyền thống, sau đó mở rộng sang học sâu và học tăng cường.

1.2.1 Phân tích và dự báo sử dụng học máy truyền thống

Trong giai đoạn đầu của việc ứng dụng trí tuệ nhân tạo vào giao thông, các kỹ thuật học máy truyền thống đóng vai trò quan trọng trong việc chuyển đổi từ điều khiển thụ động sang điều khiển có tính dự báo. Các phương pháp này thường được sử dụng để xử lý tính bất định của trạng thái giao thông hoặc dự báo lưu lượng trong ngắn hạn, từ đó cung cấp thông tin hỗ trợ cho các quyết định điều khiển.

1.2.1.1 Hệ thống điều khiển dựa trên logic mờ

Logic mờ là một trong những kỹ thuật sớm được áp dụng để xử lý tính bất định và mập mờ của các tham số giao thông như "mật độ cao" hay "hàng đợi dài". Các bộ điều khiển mờ mô phỏng kinh nghiệm của con người trong việc điều chỉnh thời gian xanh thông qua các luật dạng "Nếu – Thì". Ưu điểm lớn của mô hình mờ là hệ thống không cần một mô hình toán học chính xác của luồng xe, nhờ đó có thể vận hành linh hoạt hơn so với điều khiển chu kỳ cố định. Tuy nhiên, việc thiết lập các hàm thành viên và tập luật mờ thường phụ thuộc nhiều vào kiến thức chuyên gia. Khi số lượng nút giao tăng lên, quá trình thiết kế và tối ưu hóa tập luật trở nên phức tạp hơn đáng kể.

1.2.1.2 Dự báo lưu lượng với học máy giám sát

Các thuật toán học máy giám sát như hồi quy vectơ hỗ trợ (*Support Vector Regression* - SVR), thuật toán K lân cận gần nhất (*K-Nearest Neighbors* - KNN) và rừng ngẫu nhiên thường được sử dụng để dự báo lưu lượng giao thông trong ngắn hạn, chẳng hạn trong khoảng từ 5 đến 15 phút tiếp theo. Với SVR, mô hình sử dụng hàm hạt nhân để biểu diễn quan hệ phi tuyến giữa dữ liệu lịch sử và lưu lượng giao thông cần dự báo. Trong điều kiện giao thông tương đối ổn định, phương pháp này có thể cho kết quả dự báo tốt, nhưng hiệu quả của mô hình phụ thuộc đáng kể vào cách lựa chọn đặc trưng, hàm hạt nhân và tham số huấn luyện. Trong khi đó, KNN dựa trên giả định rằng các trạng thái giao thông tương tự trong quá khứ có xu hướng lặp lại trong hiện tại. Phương pháp này có ưu điểm là không cần tham số hóa phức tạp và có thể thích nghi với các mẫu dữ liệu mới, nhưng chi phí tính toán tăng nhanh khi tập dữ liệu mẫu trở nên lớn.

1.2.1.3 Hạn chế của học máy truyền thống trong điều khiển giao thông

Mặc dù các kỹ thuật học máy truyền thống mở rộng đáng kể khả năng phân tích và dự báo so với các phương pháp cổ điển, chúng vẫn tồn tại một số hạn chế khi được sử dụng cho bài toán điều khiển giao thông. Trước hết, nhiều mô hình trong nhóm này chủ yếu dựa trên các đặc trưng được thiết kế thủ công, nên khó phản ánh đầy đủ những biến động mang tính hệ thống của luồng xe hỗn hợp. Bên cạnh đó, quy trình điều khiển thường được tổ chức thành hai bước tách biệt, gồm dự báo trạng thái giao thông trước rồi mới đưa ra quyết định điều khiển. Cách tổ chức này có thể tạo ra độ trễ nhất định và làm giảm khả năng phản ứng của hệ thống trước các sự cố ùn tắc đột ngột. Ngoài ra, phần lớn các nghiên cứu trong giai đoạn này tập trung tối ưu cho từng

nút giao đơn lẻ, trong khi cơ chế chia sẻ thông tin và phối hợp trên toàn mạng lưới vẫn còn hạn chế.

1.2.2 Nhận biết trạng thái giao thông với học sâu

Sự phát triển của thị giác máy tính cho phép sử dụng mạng nơ-ron tích chập (*Convolutional Neural Network* - CNN) và các kiến trúc tuần tự như mạng nơ-ron tái hồi (*Recurrent Neural Network* - RNN) hoặc mạng bộ nhớ dài-ngắn hạn (*Long Short-Term Memory* - LSTM) để trích xuất đặc trưng giao thông từ luồng dữ liệu camera. Cách tiếp cận này chủ yếu đóng vai trò nhận biết và biểu diễn trạng thái giao thông, chẳng hạn như nhận diện mật độ phương tiện, phân loại xe máy và ô tô, hoặc ước lượng diễn biến hàng đợi theo thời gian. Các đặc trưng thu được từ mô hình học sâu có thể cung cấp đầu vào giàu thông tin hơn cho hệ thống điều khiển phía sau. Tuy nhiên, bản thân mô hình nhận biết trạng thái thường chưa trực tiếp quyết định pha đèn, mà cần được kết hợp với một cơ chế điều khiển hoặc tối ưu hóa khác.

1.2.3 Điều khiển trực tiếp với học tăng cường

Học tăng cường tạo ra bước chuyển quan trọng từ cách tiếp cận dự báo rồi mới điều khiển sang cách học chính sách điều khiển trực tiếp thông qua tương tác với môi trường. Sự kết hợp giữa học tăng cường và mạng nơ-ron sâu, tiêu biểu là kiến trúc mạng Q sâu (*Deep Q-Network* - DQN) [5], đã mở ra khả năng xử lý các không gian trạng thái phức tạp mà các phương pháp truyền thống khó tiếp cận. Trong lĩnh vực điều khiển tín hiệu giao thông, các nghiên cứu như IntelliLight [6] cho thấy tiềm năng của việc học chính sách điều khiển thông qua thử và sai. Sau đó, các mô hình như phương pháp tối ưu pha đèn (*Fair Resource Allocation for Phases* - FRAP) [7] tiếp tục cải tiến hướng tiếp cận này bằng cách mô hình hóa sự cạnh tranh giữa các pha đèn và bảo đảm tính bất biến đối với cấu trúc nút giao.

Để có cái nhìn tổng quan về sự phát triển của các mô hình học tăng cường trong giao thông, Bảng 1.1 tóm tắt đặc trưng và hạn chế của một số mô hình tiêu biểu.

Bảng 1.1: So sánh các mô hình học tăng cường sâu trong điều khiển đèn tín hiệu

Mô hình	**Thuật toán**	**Đặc trưng chính**	**Hạn chế**
IntelliLight [6]	DQN	Học đa mục tiêu (phần thưởng phức hợp)	Chỉ tối ưu cho một nút giao đơn lẻ.
PressLight [4]	DQN	Tích hợp lý thuyết Max-Pressure	Khó phối hợp trong mạng lưới không đối xứng.
CoLight [8]	Chú ý đồ thị	Cơ chế chú ý giữa các nút lân cận	Độ phức tạp tính toán cao khi mạng lưới lớn.
FRAP [7]	DQN	Bất biến với cấu trúc nút giao	Chưa giải quyết tốt vấn đề phối hợp mạng lưới.
MPLight [9]	FRAP + áp lực	Khả năng mở rộng quy mô mạng	Đòi hỏi dữ liệu hàng đợi chính xác ở các làn kế tiếp.

1.2.4 Học tăng cường và PPO

Các mô hình DQN kể trên hoạt động hiệu quả trong không gian hành động rời rạc và tương đối nhỏ, nhưng chúng gặp khó khăn khi mở rộng do phải ước lượng giá trị cho từng hành động hợp lệ tại mỗi trạng thái. Trong khi đó, nhóm phương pháp tối ưu hóa chính sách giải quyết bài toán điều khiển trực tiếp bằng cách học một chính sách $\pi(a|s)$ ánh xạ trạng thái sang xác suất chọn hành động. Cách tiếp cận này phù hợp hơn với các bài toán có không gian trạng thái lớn và yêu cầu quá trình cập nhật chính sách ổn định.

Trong số các thuật toán tối ưu hóa chính sách, tối ưu hóa chính sách cận gần (*Proximal Policy Optimization* - PPO) [10] là một lựa chọn phổ biến nhờ tính ổn định và tương đối dễ triển khai. PPO cải thiện thuật toán REINFORCE [11] và các phương pháp gradient chính sách cơ bản bằng cách sử dụng hàm mục tiêu có cắt ngưỡng. Cơ chế cắt ngưỡng giới hạn mức thay đổi của chính sách trong mỗi bước cập nhật, qua đó giảm nguy cơ làm mất ổn định quá trình học do các bước cập nhật quá lớn. Cơ chế này cũng cho phép PPO tận dụng nhiều vòng cập nhật từ cùng một lô dữ liệu, giúp tăng hiệu quả sử dụng mẫu so với các phương pháp gradient chính sách truyền thống. Ngoài ra, PPO thường được kết hợp với ước lượng lợi thế GAE (ước lượng lợi thế tổng quát (*Generalized Advantage Estimation* - GAE)) để giảm phương sai khi đánh

giá chất lượng hành động. Những đặc điểm này làm cho PPO trở thành một lựa chọn phù hợp cho các bài toán điều khiển có không gian trạng thái lớn và yêu cầu học ổn định, bao gồm điều khiển tín hiệu giao thông [4, 9].

1.2.5 Sự phát triển của cơ chế phối hợp và quy mô mạng lưới

Theo các khảo sát gần đây [12], xu hướng nghiên cứu trong điều khiển tín hiệu giao thông đang dịch chuyển từ tối ưu hóa một nút giao đơn lẻ sang điều khiển phối hợp trên toàn mạng lưới. Các giải pháp như CoLight [8] sử dụng cơ chế chú ý để hỗ trợ các nút giao trao đổi thông tin trạng thái với nhau, từ đó góp phần giảm hiện tượng ùn tắc lan truyền. Việc kết hợp học tăng cường sâu với các mô hình biểu diễn không gian vì vậy trở thành một hướng tiếp cận quan trọng trong điều khiển giao thông đô thị.

1.3 Học tăng cường đa tác tử trong giao thông

Khi bài toán điều khiển tín hiệu giao thông được mở rộng từ một bộ điều khiển đèn tín hiệu đơn lẻ sang một mạng lưới gồm nhiều nút giao có liên kết chặt chẽ, việc sử dụng một tác tử duy nhất để điều khiển toàn bộ hệ thống trở nên kém khả thi. Nguyên nhân là không gian trạng thái và không gian hành động tăng nhanh theo số lượng nút giao, trong khi quyết định tại một nút có thể ảnh hưởng trực tiếp đến các nút lân cận. Trong bối cảnh đó, học tăng cường đa tác tử cung cấp một cách tiếp cận phân tán tự nhiên, trong đó mỗi bộ điều khiển có thể được mô hình hóa như một tác tử riêng và cùng tham gia vào quá trình tối ưu hóa vận hành của mạng lưới.

1.3.1 Khung lý thuyết Markov đa tác tử

Trong học tăng cường đa tác tử, bài toán điều khiển tín hiệu giao thông thường được mô hình hóa dưới dạng trò chơi Markov [13]. Mô hình này cho phép mô tả quá trình nhiều tác tử cùng quan sát môi trường, lựa chọn hành động và nhận phần thưởng sau mỗi bước tương tác. Mục 2.3.1 trình bày chi tiết cách đề án đặc tả trạng thái, hành động và phần thưởng cho bài toán này.

Trong phạm vi đề án, một tác tử $i \in \mathcal{N}$ tương ứng với một bộ điều khiển đèn tín hiệu có nhiều hơn một giai đoạn xanh hợp lệ. Tại mỗi bước thời gian, tác tử quan sát trạng thái cục bộ s_i, thực hiện hành động a_i và nhận phần thưởng r_i dựa trên hiệu quả điều khiển tại nút giao cũng như ảnh hưởng của quyết định đó đến luồng giao thông chung.

1.3.2 Thách thức trong điều khiển đa tác tử

Việc áp dụng học tăng cường đa tác tử vào điều khiển giao thông phải đối mặt với nhiều thách thức. Thách thức đầu tiên là tính không dừng của môi trường học. Do các tác tử cùng học tập và thay đổi chính sách trong quá trình huấn luyện, môi trường quan sát được từ góc nhìn của mỗi tác tử không còn ổn định theo thời gian. Trong mạng lưới giao thông, hành động của một bộ điều khiển có thể làm thay đổi dòng xe đi vào các nút giao lân cận, khiến việc tối ưu hóa chính sách trở nên khó khăn hơn [14].

Thách thức thứ hai là khả năng mở rộng. Khi số lượng nút giao tăng lên, việc phối hợp thông tin giữa các tác tử đòi hỏi nhiều tài nguyên tính toán và một cơ chế biểu diễn phù hợp để mô hình hóa quan hệ phụ thuộc giữa các nút. Nếu mỗi tác tử chỉ tối ưu quyết định cục bộ mà không xét đến tác động lên các nút lân cận, hệ thống có thể giảm hiệu quả trên quy mô toàn mạng, đặc biệt trong các khu vực có khoảng cách giữa các nút giao ngắn và hiện tượng ùn tắc lan truyền mạnh.

1.3.3 Chiến lược phối hợp

Để giải quyết vấn đề phối hợp trong hệ thống đa tác tử, các nghiên cứu thường xem xét hai hướng tiếp cận chính. Hướng thứ nhất là học độc lập, trong đó mỗi tác tử tự học chính sách của mình và coi các tác tử khác như một phần của môi trường. Cách tiếp cận này có ưu điểm là đơn giản và dễ triển khai, nhưng nó thường gặp khó khăn trong các môi trường phức tạp do sự thay đổi đồng thời của nhiều chính sách có thể làm quá trình học trở nên không ổn định.

Hướng thứ hai là huấn luyện tập trung và thực thi phân tán. Trong khung này, các tác tử có thể sử dụng thông tin toàn cục trong giai đoạn huấn luyện, nhưng khi triển khai, mỗi tác tử chỉ dựa trên quan sát cục bộ để ra quyết định. Các kiến trúc như thuật toán điều khiển đa tác tử (*Multi-Agent Deep Deterministic Policy Gradient* - MADDPG) [15] đã đặt nền móng cho việc sử dụng bộ phê bình tập trung nhằm cải thiện hiệu quả học. Bên cạnh đó, các mô hình phân rã giá trị như phân rã giá trị QMIX (*Q-Mixer Value Decomposition* - QMIX) [16] và các phương pháp tối ưu hóa chính sách như PPO đa tác tử (*Multi-Agent PPO* - MAPPO) [17] cho thấy hiệu quả trong việc xử lý bài toán phối hợp và phân bổ tín dụng trong các hệ thống quy mô lớn.

Gần đây, tối ưu hóa chính sách tác tử không đồng nhất (*Heterogeneous-Agent PPO* - HAPPO) [18] mở rộng MAPPO theo hướng cho phép mỗi tác tử sở hữu một chính sách riêng thay vì dùng chung chính sách. Cách tiếp cận này phù hợp với các hệ thống có tác tử không đồng nhất, chẳng hạn các nút giao có cấu trúc làn, số pha và vai trò

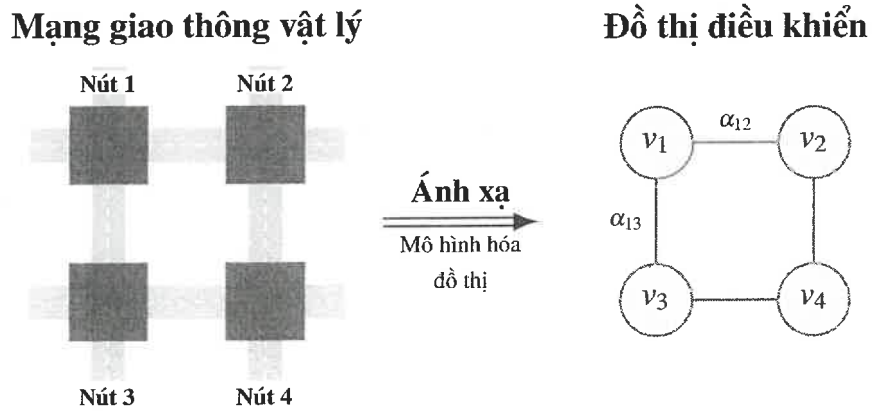
giao thông khác nhau trong cùng một mạng lưới. HAPPO sử dụng cơ chế cập nhật tuần tự với tỷ lệ quan trọng tích lũy để tăng tính ổn định khi số lượng tác tử tăng, đồng thời vẫn giữ được lợi ích của khung huấn luyện tập trung và thực thi phân tán.

1.4 Mạng nơ-ron đồ thị cho bài toán phối hợp

Mạng lưới giao thông có cấu trúc tự nhiên dưới dạng đồ thị, trong đó các bộ điều khiển tại nút giao có thể được xem là các đỉnh và các đoạn đường kết nối giữa các nút giao có thể được xem là các cạnh. Do đó, mạng nơ-ron đồ thị là một công cụ phù hợp để trích xuất đặc trưng không gian và mô hình hóa sự tương tác giữa các nút giao trong mạng lưới.

1.4.1 Biểu diễn đồ thị giao thông

Thay vì coi mỗi bộ điều khiển đèn tín hiệu là một thực thể độc lập, mạng nơ-ron đồ thị cho phép mô hình biểu diễn các quan hệ phụ thuộc giữa các nút giao. Trong cách biểu diễn này, trạng thái của một bộ điều khiển không chỉ phản ánh hàng đợi hoặc mật độ phương tiện tại nút giao đó, mà còn có thể tích hợp thông tin từ các nút lân cận thông qua cấu trúc đồ thị [19]. Các đặc trưng được sử dụng trong biểu diễn đồ thị thường bao gồm mật độ phương tiện, độ dài hàng đợi trung bình và tốc độ luồng xe trên các cạnh kết nối. Nhờ đó, mô hình có thể tiếp cận một mô tả đầy đủ hơn về áp lực giao thông cục bộ và ảnh hưởng không gian giữa các nút giao.



* **Ghi chú:** Mỗi đỉnh là một bộ điều khiển đèn tín hiệu, cạnh biểu diễn kết nối trực tiếp giữa các bộ điều khiển. Trong biến thể GAT, mỗi luồng truyền tin từ đỉnh v_j đến đỉnh v_i được gán trọng số chú ý α_{ij}.

Hình 1.3: Minh họa ánh xạ mạng giao thông thành đồ thị điều khiển

1.4.2 Cơ chế truyền tin và sự chú ý

Cơ chế cốt lõi của mạng nơ-ron đồ thị là quá trình truyền tin giữa các nút. Trong bài toán điều khiển tín hiệu giao thông, mỗi nút trong đồ thị có thể biểu diễn một tác tử điều khiển, còn các cạnh thể hiện quan hệ lân cận hoặc khả năng ảnh hưởng giữa các nút giao. Thông qua quá trình truyền tin, một tác tử không chỉ sử dụng thông tin về hàng đợi cục bộ tại nút giao của mình, mà còn có thể xét đến trạng thái của các nút lân cận có khả năng làm thay đổi dòng xe đến hoặc dòng xe thoát của nó.

Khác với mạng tích chập đồ thị (*Graph Convolutional Network* - GCN), trong đó thông tin từ các nút lân cận thường được tổng hợp theo cấu trúc cố định của ma trận kề, mạng chú ý đồ thị (*Graph Attention Network* - GAT) học trọng số chú ý α_{ij} cho từng cặp nút lân cận. Trọng số này cho phép mô hình gán mức ảnh hưởng khác nhau cho các tác tử lân cận tùy theo trạng thái hiện thời. Ví dụ, một nút giao đang có hàng đợi lớn hoặc áp lực cao có thể được gán trọng số lớn hơn so với một nút giao đang thông thoáng. Cách tổng hợp có trọng số này giúp mô hình tập trung vào các quan hệ giao thông quan trọng, thay vì xem mọi nút lân cận là như nhau. Các nghiên cứu như CoLight [8] đã khai thác ý tưởng này để hỗ trợ phối hợp giữa các nút giao trong mạng lưới.

1.4.3 Khả năng mở rộng và điều khiển mạng lưới lớn

Các tiếp cận như MPLight [9] đã kết hợp nguyên lý áp lực cực đại với học tăng cường sâu để xử lý bài toán giao thông trên mạng nhiều nút giao. Việc tích hợp đặc trưng từ các mạng tích chập đồ thị như GCN [20] hoặc mạng chú ý đồ thị như GAT [21] giúp mô hình khai thác tốt hơn cấu trúc không gian của mạng lưới đường bộ. Ngoài ra, các kiến trúc đồ thị không-thời gian như mạng tích chập đồ thị không-thời gian (*Spatio-Temporal Graph Convolutional Network* - STGCN) [22] cũng mở ra hướng nghiên cứu mới trong việc đồng thời nắm bắt đặc trưng không gian của mạng lưới và đặc trưng thời gian của luồng phương tiện. Những hướng tiếp cận này cho thấy vai trò quan trọng của biểu diễn đồ thị trong các bài toán điều khiển giao thông đô thị có quy mô lớn và có sự phụ thuộc chặt chẽ giữa các nút giao.

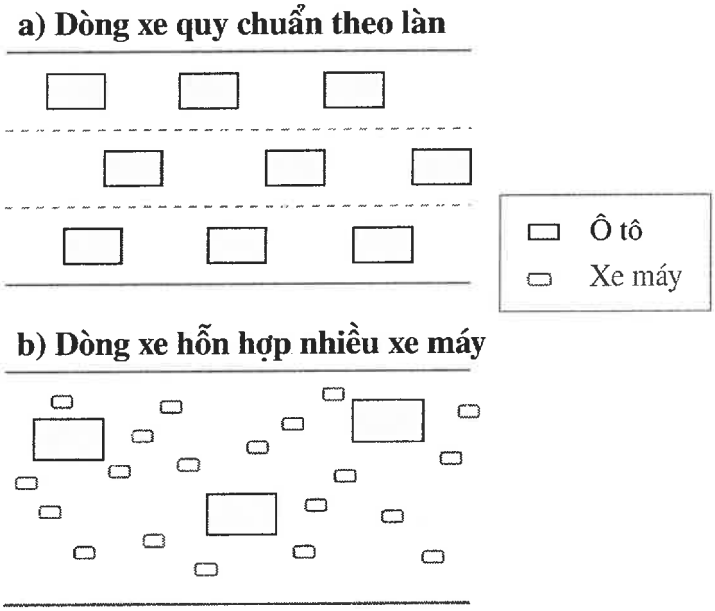
1.5 Phân tích khoảng trống nghiên cứu

Mặc dù các nghiên cứu quốc tế đã đạt được nhiều tiến bộ trong điều khiển tín hiệu giao thông, việc áp dụng trực tiếp các mô hình này vào bối cảnh Việt Nam vẫn gặp nhiều rào cản. Những rào cản này xuất phát từ đặc thù dòng xe hỗn hợp, hạn chế về dữ liệu giao thông chuẩn hóa và yêu cầu phối hợp giữa các nút giao trong mạng lưới

đô thị dày đặc. Vì vậy, việc đánh giá các phương pháp điều khiển hiện đại trên một kịch bản mô phỏng phản ánh tốt hơn điều kiện giao thông Việt Nam là cần thiết.

1.5.1 Sự thích nghi với luồng giao thông hỗn hợp

Phần lớn các nghiên cứu hiện có dựa trên giả định rằng phương tiện di chuyển theo làn đường tương đối ổn định. Tuy nhiên, tại các đô thị lớn như Hà Nội và Thành phố Hồ Chí Minh, dòng giao thông có tính hỗn hợp cao với tỷ lệ xe máy lớn. Xe máy thường có hành vi lấn làn và tận dụng các khoảng trống giữa các phương tiện, khiến các mô hình đếm xe hoặc ước lượng hàng đợi theo làn truyền thống trở nên kém chính xác hơn. Vì vậy, một hệ thống điều khiển tín hiệu giao thông phù hợp với bối cảnh Việt Nam cần có khả năng phản ánh đặc điểm của dòng xe hỗn hợp, thay vì chỉ dựa trên giả định về luồng xe quy chuẩn theo làn.



Hình 1.4: So sánh dòng xe quy chuẩn theo làn và dòng xe hỗn hợp thực tế tại Việt Nam

1.5.2 Hạ tầng dữ liệu và môi trường mô phỏng

Các bộ dữ liệu chuẩn quốc tế chưa phản ánh đầy đủ thực trạng giao thông Việt Nam, đặc biệt là đặc trưng xe máy chiếm tỷ lệ cao và hành vi di chuyển linh hoạt trong không gian đường. Để giảm khoảng cách này, đề án sử dụng dữ liệu khảo sát giao thông từ dự án UTM cho khu vực Hà Nội để sinh nhu cầu giao thông và xây dựng kịch bản mô phỏng vi mô tại khu vực Thanh Xuân. Mạng lưới đường bộ được lấy từ OSM và được xử lý để phù hợp với mô phỏng SUMO. Cách tiếp cận này cho phép

quá trình huấn luyện và đánh giá các tác tử điều khiển được thực hiện trong một môi trường gần với bối cảnh nghiên cứu hơn so với các mạng lưới tổng hợp, đồng thời vẫn thừa nhận các hạn chế của nguồn dữ liệu đầu vào.

1.5.3 Tính phối hợp giữa các nút giao trong mạng lưới dày đặc

Trong các đô thị Việt Nam, nhiều nút giao có khoảng cách ngắn và chịu ảnh hưởng lẫn nhau mạnh. Quyết định điều khiển tại một nút có thể làm thay đổi dòng xe đến của các nút lân cận, từ đó tạo ra hiện tượng ùn tắc lan truyền nếu các bộ điều khiển hoạt động hoàn toàn độc lập. Vì vậy, bài toán điều khiển tín hiệu giao thông trên mạng lưới đô thị không chỉ yêu cầu tối ưu trạng thái cục bộ tại từng nút, mà còn cần xem xét cơ chế phối hợp giữa các tác tử điều khiển.

Từ các khoảng trống trên, đề án tập trung đánh giá có hệ thống các thuật toán học tăng cường đa tác tử dựa trên PPO trong bối cảnh giao thông hỗn hợp tại Việt Nam. Cụ thể, đề án triển khai và so sánh MAPPO, HAPPO, GAT-MAPPO và GAT-HAPPO trên kịch bản Thanh Xuân UTM, trong đó các biến thể GAT khai thác quan hệ lân cận giữa các nút giao thông qua đồ thị chính sách. Đóng góp của đề án vì vậy nằm ở việc xây dựng một môi trường mô phỏng vi mô có xét đến dữ liệu khảo sát thực tế, đặc tả bài toán điều khiển tín hiệu giao thông dưới dạng đa tác tử, và cung cấp kết quả đánh giá thực nghiệm giữa các biến thể MARL cùng với các phương pháp đối chứng. Đề án không đề xuất một thuật toán lý thuyết mới, mà tập trung vào đóng góp ứng dụng và thực nghiệm trong bối cảnh giao thông đô thị Việt Nam.

1.6 Kết luận chương

Chương này đã trình bày các nền tảng và hướng nghiên cứu liên quan đến điều khiển tín hiệu giao thông, bắt đầu từ các phương pháp truyền thống dựa trên mô hình toán học đến các tiếp cận hiện đại dựa trên học tăng cường, học tăng cường đa tác tử và mạng nơ-ron đồ thị. Từ phần tổng quan, có thể thấy Max-Pressure là một phương pháp đối chứng thích ứng quan trọng cần được xem xét trong quá trình đánh giá, trong khi MAPPO và HAPPO là các thuật toán MARL phù hợp để nghiên cứu bài toán phối hợp giữa nhiều bộ điều khiển. Bên cạnh đó, cơ chế biểu diễn đồ thị và truyền tin giữa các nút giao là một hướng tiếp cận phù hợp khi mở rộng điều khiển từ một nút đơn lẻ sang mạng lưới đô thị.

Trên cơ sở các khoảng trống đã phân tích, chương tiếp theo trình bày mô hình học tăng cường đa tác tử cho bài toán điều khiển tín hiệu giao thông. Nội dung chương tiếp theo bao gồm cách xây dựng môi trường mô phỏng, đặc tả tác tử điều khiển, định

nghĩa không gian quan sát, không gian hành động và hàm phần thưởng, cũng như kiến trúc của các thuật toán MAPPO, HAPPO, GAT-MAPPO và GAT-HAPPO được triển khai trong đề án.

Chương 2
MÔ HÌNH HỌC TĂNG CƯỜNG ĐA TÁC TỬ TRONG ĐIỀU KHIỂN TÍN HIỆU GIAO THÔNG

Chương này trình bày mô hình học tăng cường đa tác tử cho bài toán điều khiển đèn tín hiệu giao thông. Nội dung chính của chương bao gồm nền tảng học tăng cường đa tác tử, cơ chế cập nhật chính sách, cách mô hình hóa bài toán, thiết kế môi trường mô phỏng, xây dựng đồ thị chính sách, kiến trúc của bốn thuật toán MARL được triển khai và giao thức huấn luyện nhằm bảo đảm tính tái lập.

2.1 Nền tảng học tăng cường đa tác tử

2.1.1 Từ MDP đến trò chơi Markov

Trong học tăng cường đơn tác tử, bài toán thường được mô hình hóa dưới dạng quá trình quyết định Markov (*Markov Decision Process* - MDP). Khi có nhiều tác tử cùng tương tác với môi trường, mô hình này được mở rộng một cách tự nhiên thành trò chơi Markov [13]. Trong bài toán nghiên cứu, mỗi tác tử đóng vai trò là bộ điều khiển đèn tín hiệu tại một nút giao, trong đó hành động của một tác tử có thể ảnh hưởng trực tiếp đến trạng thái của các nút giao lân cận.

Khi quan sát của từng tác tử chỉ giới hạn trong phạm vi cục bộ, bài toán trở thành một dạng bài toán nhiều tác tử với khả năng quan sát không đầy đủ. Điều này phản ánh sát điều kiện thực tế của hệ thống điều khiển đèn tín hiệu, khi mỗi nút giao chỉ quan sát được thông tin về các làn vào/ra, pha tín hiệu hiện tại và thời gian đã trôi qua của chính nó, thay vì toàn bộ trạng thái của mạng lưới giao thông.

2.1.2 Huấn luyện tập trung, thực thi phân tán

Khung huấn luyện tập trung - thực thi phân tán (*Centralized Training - Decentralized Execution* - CTDE) phù hợp với bài toán điều khiển tín hiệu giao thông đa nút giao. Trong giai đoạn huấn luyện, mạng giá trị tập trung hoặc bộ mã hóa đồ thị có thể khai thác thông tin toàn cục nhằm giảm phương sai trong quá trình ước lượng giá trị. Trong giai đoạn thực thi, mỗi tác tử chỉ sử dụng chính sách cục bộ của mình để lựa chọn hành động. Cách tiếp cận này vừa bảo đảm tính khả thi khi triển khai, vừa tạo điều kiện để các tác tử học được cơ chế phối hợp giữa các nút giao.

2.1.3 *Các thách thức chính trong MARL*

Việc áp dụng MARL vào bài toán điều khiển đèn tín hiệu đối mặt với một số thách thức chính. Thách thức đầu tiên là tính không dừng của môi trường học. Do chính sách của các tác tử khác liên tục thay đổi trong quá trình huấn luyện, động lực của môi trường quan sát được bởi mỗi tác tử cũng biến đổi theo thời gian. Điều này làm cho quá trình tối ưu hóa chính sách trở nên khó khăn hơn so với trường hợp đơn tác tử. Thách thức thứ hai là việc xác định mức độ đóng góp của từng tác tử đối với kết quả chung. Trong hệ thống nhiều nút giao, một phần thưởng chung hoặc phần thưởng cục bộ có thể chưa phản ánh đầy đủ vai trò của từng tác tử trong trạng thái vận hành cuối cùng của mạng lưới. Bên cạnh đó, khả năng mở rộng cũng là một vấn đề quan trọng, vì không gian hành động chung tăng nhanh khi số lượng nút giao tăng lên. Ngoài ra, bài toán điều khiển đèn tín hiệu còn có đặc thù riêng là không phải mọi hành động đều hợp lệ tại mọi thời điểm, do các quyết định chuyển pha phải tuân thủ chu kỳ đèn và ràng buộc an toàn.

Để giảm thiểu các khó khăn trên, đề án áp dụng các kỹ thuật như chia sẻ tham số, bộ lọc hành động và biểu diễn đồ thị. Chia sẻ tham số giúp giảm số lượng tham số cần học trong các mô hình dùng chính sách chung. Bộ lọc hành động bảo đảm tác tử chỉ lựa chọn các hành động hợp lệ tại thời điểm ra quyết định. Biểu diễn đồ thị cho phép mô hình khai thác quan hệ lân cận giữa các nút giao, qua đó hỗ trợ quá trình phối hợp trong mạng lưới giao thông.

2.2 Cơ chế cập nhật chính sách điều khiển

2.2.1 *Giới hạn mức thay đổi bằng PPO*

Trong quá trình huấn luyện, chính sách điều khiển cần được cập nhật một cách ổn định. Nếu chính sách thay đổi quá lớn sau một lần cập nhật, mô hình có thể suy giảm hiệu năng hoặc học theo hướng không ổn định. PPO được sử dụng để kiểm soát mức độ thay đổi của chính sách trong mỗi bước tối ưu. Trước hết, thuật toán so sánh xác suất lựa chọn cùng một hành động theo chính sách mới và chính sách cũ thông qua tỷ lệ sau:

$$\rho_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}. \tag{2.1}$$

Trong đó, $\pi_{\theta_{\text{old}}}$ là chính sách trước khi cập nhật, còn π_θ là chính sách sau khi cập nhật. Để ngăn chính sách mới lệch quá xa so với chính sách cũ, PPO sử dụng hàm mục tiêu cắt ngưỡng:

$$L^{CLIP}(\theta) = \mathbb{E}\left[\min\left(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right]. \tag{2.2}$$

Cơ chế cắt ngưỡng này giúp quá trình học diễn ra ổn định hơn vì chính sách không bị thay đổi quá mạnh chỉ dựa trên một lô dữ liệu mô phỏng. Trong đề án, PPO đóng vai trò là cơ chế cập nhật chính sách nền cho cả bốn thuật toán MAPPO, HAPPO, GAT-MAPPO và GAT-HAPPO.

2.2.2 Ước lượng lợi thế bằng GAE

Sau mỗi lượt tương tác với môi trường mô phỏng, mô hình cần xác định hành động đã thực hiện tốt hơn hay kém hơn mức kỳ vọng. Đại lượng dùng để phản ánh mức chênh lệch này được gọi là lợi thế, ký hiệu là $\hat{A}_t$. Trong đề án, lợi thế được ước lượng bằng GAE từ chuỗi phần thưởng và giá trị trạng thái do critic dự đoán.

Trước hết, sai số dự báo giá trị tại thời điểm t được tính như sau:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t), \tag{2.3}$$

trong đó r_t là phần thưởng do môi trường trả về sau khi tác tử thực hiện hành động, $V(s_t)$ là giá trị trạng thái hiện tại do critic ước lượng, còn $V(s_{t+1})$ là giá trị của trạng thái kế tiếp. Đại lượng δ_t cho biết giá trị thực tế quan sát được sau một bước chuyển trạng thái lệch bao nhiêu so với giá trị mà critic dự đoán.

Tiếp theo, GAE tổng hợp các sai số này theo thời gian để tạo ra ước lượng lợi thế:

$$\hat{A}_t^{GAE} = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}. \tag{2.4}$$

Tham số λ điều chỉnh mức độ khai thác thông tin dài hạn trong quá trình ước lượng. Khi λ nhỏ, ước lượng thường ổn định hơn nhưng ít phản ánh ảnh hưởng dài hạn của hành động. Khi λ tiến gần đến 1, ước lượng sử dụng nhiều thông tin tương lai hơn nhưng có thể nhiễu hơn. Trong đề án này, $\lambda = 0.95$ được sử dụng như một cấu hình cố định nhằm cân bằng giữa độ ổn định và khả năng khai thác thông tin dài hạn.

Cần phân biệt giữa phần thưởng, critic và GAE. Phần thưởng r_t là tín hiệu gốc do môi trường trả về sau khi tác tử thực hiện hành động. Critic là mạng giá trị dùng để ước lượng mức độ tốt của trạng thái, ký hiệu là $V(s)$. GAE không phải là phần thưởng và cũng không phải là critic; đây là phương pháp kết hợp chuỗi phần thưởng với các giá trị $V(s_t)$ và $V(s_{t+1})$ do critic dự đoán để tính lợi thế $\hat{A}_t$. Lợi thế cho biết hành

động vừa thực hiện mang lại kết quả tốt hơn hay kém hơn giá trị kỳ vọng do critic ước lượng, từ đó được PPO sử dụng để cập nhật actor.

2.2.3 Phương pháp sử dụng PPO và GAE trong quá trình huấn luyện

Trong quy trình huấn luyện của đề án, actor, critic, phần thưởng, GAE và PPO đảm nhiệm các vai trò khác nhau. Actor nhận quan sát hiện tại của tác tử và sinh phân phối xác suất trên các hành động hợp lệ. Sau khi actor chọn hành động, môi trường mô phỏng cập nhật trạng thái và trả về phần thưởng tương ứng. Phần thưởng này không được actor sử dụng trực tiếp tại thời điểm ra quyết định, mà được lưu lại trong dữ liệu huấn luyện.

Critic sử dụng trạng thái để ước lượng giá trị kỳ vọng của trạng thái đó. GAE sau đó kết hợp chuỗi phần thưởng với các giá trị do critic dự đoán để tính lợi thế $\hat{A}_t$. Lợi thế cho biết hành động đã thực hiện mang lại kết quả tốt hơn hay kém hơn mức kỳ vọng. Cuối cùng, PPO sử dụng lợi thế này để cập nhật actor theo hướng tăng xác suất của các hành động có lợi thế dương và giảm xác suất của các hành động có lợi thế âm, đồng thời vẫn giới hạn mức thay đổi của chính sách thông qua hàm mục tiêu cắt ngưỡng.

Như vậy, trong đề án, GAE không phải là phần thưởng và cũng không phải là critic. GAE là phương pháp tạo tín hiệu lợi thế từ phần thưởng và giá trị do critic ước lượng. PPO sử dụng tín hiệu lợi thế đó để cập nhật chính sách một cách ổn định. Hai thành phần này tạo thành cơ chế huấn luyện chung cho các thuật toán MAPPO, HAPPO, GAT-MAPPO và GAT-HAPPO được trình bày trong các phần tiếp theo.

2.3 Mô hình hóa bài toán

2.3.1 Trò chơi Markov

Dựa trên môi trường mô phỏng SUMO được xây dựng cho kịch bản Thanh Xuân UTM, đề án mô hình hóa bài toán điều khiển đèn tín hiệu tại nhiều nút giao dưới dạng trò chơi Markov. Trong môi trường thực nghiệm, mỗi tác tử đưa ra quyết định điều khiển sau mỗi 5 giây mô phỏng. Tại mỗi bước thời gian, tác tử quan sát trạng thái giao thông trong phạm vi nút giao của mình và lựa chọn một hành động điều khiển. Sau đó, môi trường mô phỏng cập nhật trạng thái giao thông và trả về phần thưởng tương ứng cho tác tử. Mô hình tổng quát được biểu diễn như sau:

$$\mathscr{M} = \langle \mathscr{N}, \mathscr{S}, \mathscr{A}, \mathscr{P}, \mathscr{R}, \gamma \rangle. \tag{2.5}$$

Trong đó:

$\mathcal{N}$: Tập tác tử điều khiển, tương ứng với các bộ điều khiển đèn tín hiệu có nhiều hơn một giai đoạn xanh.

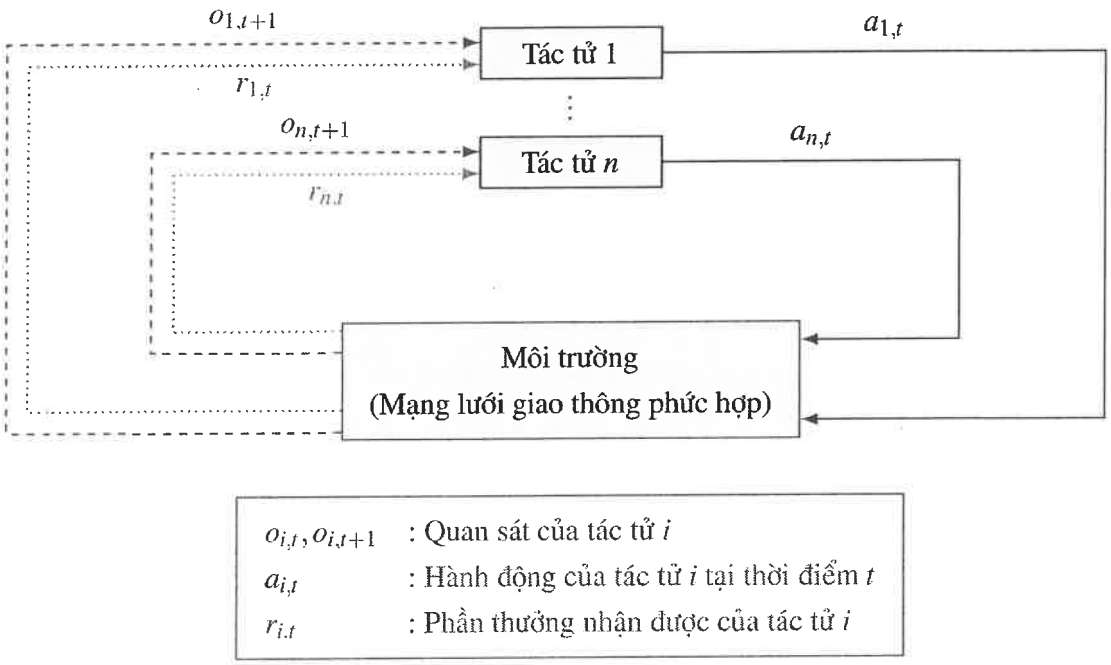
$\mathcal{S}$: Không gian trạng thái của môi trường.

$\mathcal{A}$: Không gian hành động của các tác tử.

$\mathcal{P}$: Hàm chuyển trạng thái được xác định bởi môi trường mô phỏng SUMO.

$\mathcal{R}$: Hàm phần thưởng.

γ : Hệ số chiết khấu.



Hình 2.1: Sơ đồ tương tác giữa tác tử và môi trường trong mô hình trò chơi Markov

2.3.2 Không gian quan sát

Do mỗi tác tử chỉ có khả năng quan sát cục bộ, tác tử không tiếp cận trực tiếp trạng thái của toàn bộ mạng lưới giao thông. Thay vào đó, mỗi tác tử nhận một vector quan sát gồm tám thành phần, phản ánh mật độ giao thông, trạng thái tín hiệu hiện tại và khả năng chuyển pha tại nút giao mà tác tử điều khiển:

$$o_i = \left[q_i^{in}, m_i^{in}, q_i^{out}, p_i, \Delta w_i, s_i^{norm}, e_i^{norm}, c_i\right]. \tag{2.6}$$

Trong đề án này, *làn tiếp cận* là các làn dẫn phương tiện vào vùng điều khiển của nút giao, trong khi *làn thoát* là các làn nhận phương tiện sau khi phương tiện đi qua nút giao. Các thành phần của vector quan sát được định nghĩa như sau:

q_i^{in} : Tổng PCU của các phương tiện đang dừng chờ (tốc độ ≤ 0.1 m/s) trên các làn tiếp cận.

m_i^{in} : Tổng PCU của các phương tiện đang di chuyển trên các làn tiếp cận.

q_i^{out} : Tổng PCU của các phương tiện đang dừng chờ trên các làn thoát.

p_i : Áp lực cục bộ, được tính bằng $(q_i^{in} + m_i^{in}) - x_i^{out}$, trong đó x_i^{out} là tổng PCU của tất cả phương tiện trên các làn thoát.

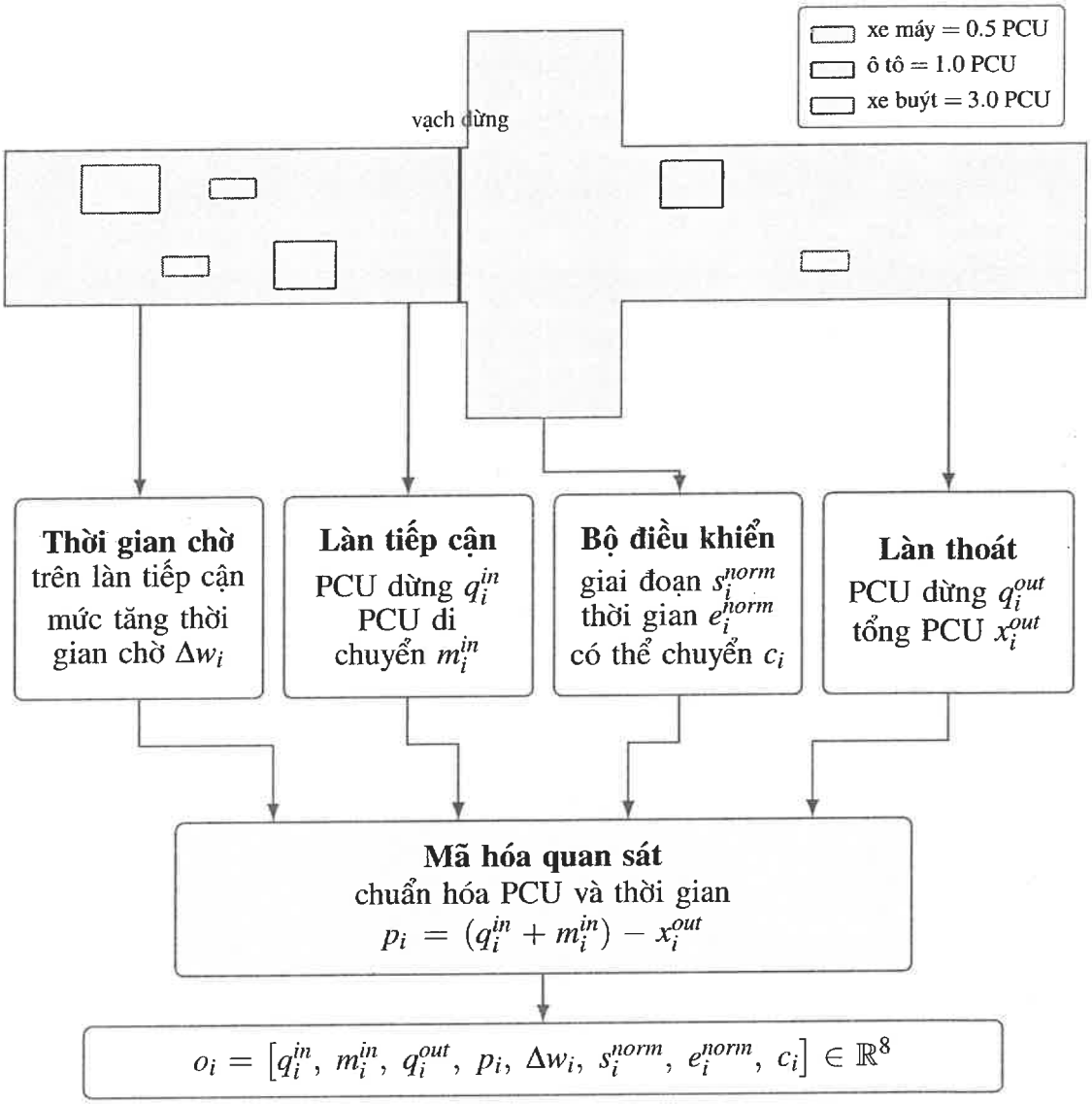
Δw_i : Tổng thời gian chờ tăng thêm của các phương tiện trên các làn tiếp cận trong khoảng thời gian Δt.

s_i^{norm} : Chỉ số giai đoạn xanh hiện tại sau khi chuẩn hóa về khoảng $[0, 1]$.

e_i^{norm} : Thời gian đã trải qua trong giai đoạn hiện tại, được chuẩn hóa về khoảng $[0, 1]$ với giá trị tham chiếu là 60 giây.

c_i : Biến chỉ báo khả năng chuyển pha (1.0 nếu đã thỏa điều kiện thời gian xanh tối thiểu, 0.0 nếu chưa thỏa).

Mặc dù q_i^{out} không xuất hiện trực tiếp trong hàm phần thưởng, đề án vẫn cung cấp thông tin này cho tác tử nhằm hỗ trợ việc ước lượng áp lực cục bộ và khả năng giải tỏa lưu lượng tại nút giao. Đối với các biến thể sử dụng GAT, cấu trúc đồ thị được trình bày chi tiết tại Mục 2.5. Trong các biến thể này, vector đặc trưng của mỗi nút được bổ sung thêm tỷ lệ giữa số giai đoạn xanh của tác tử và số giai đoạn xanh lớn nhất trong toàn mạng lưới, tạo thành vector đầu vào gồm 9 chiều.



Hình 2.2: Quy trình trích xuất và mã hóa vector quan sát cục bộ dựa trên hệ số PCU

2.3.3 Không gian hành động và cơ chế lọc hành động

Sau khi nhận vector quan sát tại thời điểm ra quyết định, tác tử lựa chọn một hành động điều khiển tương ứng với việc duy trì hoặc yêu cầu chuyển giai đoạn xanh. Tuy nhiên, tập hành động khả dụng của tác tử luôn bị ràng buộc bởi cấu trúc chu kỳ đèn tín hiệu tại nút giao. Trong đề án, hành động a_i của tác tử được định nghĩa theo hai nhóm chính. Tác tử luôn có thể chọn hành động 0 để duy trì giai đoạn xanh hiện tại. Đối với hành động $1+k$, tác tử yêu cầu chuyển sang giai đoạn xanh k; hành động này chỉ hợp lệ khi giai đoạn k là giai đoạn kế tiếp trong chu kỳ tuần hoàn và thời gian xanh tối thiểu của giai đoạn hiện tại đã được thỏa mãn.

Do chỉ một phần hành động là hợp lệ tại từng thời điểm, phân phối xác suất hành động chỉ được chuẩn hóa trên tập hành động hợp lệ:

$$\pi(a \mid o) \propto \exp(z_a(o)), \quad a \in \mathcal{A}_i^{valid}. \tag{2.7}$$

Các hành động không hợp lệ bị loại khỏi phân phối xác suất, bảo đảm rằng tác tử không thể lựa chọn các lệnh vi phạm quy tắc vận hành của chu trình đèn tín hiệu. Trong quá trình thực thi, hệ thống gán xác suất bằng 0 cho các hành động không hợp lệ trước khi lấy mẫu hành động.

2.3.4 Ràng buộc chuyển pha an toàn

Ngay cả khi tác tử phát lệnh chuyển pha, hệ thống vẫn kiểm tra các điều kiện an toàn trước khi thực thi lệnh này. Một lệnh chuyển pha chỉ được chấp nhận khi tác tử không ở trạng thái chuyển tiếp (*in_transition*), giai đoạn được yêu cầu là giai đoạn kế tiếp trong chu kỳ tuần hoàn và thời gian duy trì giai đoạn hiện tại đã vượt ngưỡng tối thiểu G_{min}.

Khi các điều kiện trên được thỏa mãn, hệ thống không chuyển trực tiếp sang giai đoạn xanh mới mà phải đi qua chuỗi pha vàng và pha đỏ toàn bộ. Cơ chế này bảo đảm tính nhất quán với quy trình vận hành thực tế của hệ thống đèn tín hiệu. Trong cấu hình thực nghiệm, thời gian xanh tối thiểu được đặt là $G_{min} = 10$ giây. Do bước quyết định là 5 giây, một giai đoạn xanh phải được duy trì ít nhất hai bước quyết định trước khi tác tử có thể chuyển pha.

2.3.5 Hàm phần thưởng

Mục tiêu điều khiển của đề án không chỉ tập trung vào một chỉ số duy nhất, mà hướng tới việc đồng thời giảm chênh lệch tải giữa làn tiếp cận và làn thoát, giảm ùn tắc cục bộ, giảm thời gian chờ và hạn chế các sự kiện dịch chuyển bắt buộc. Trong SUMO, dịch chuyển bắt buộc là cơ chế tự động di chuyển phương tiện khỏi mạng lưới khi phương tiện bị mắc kẹt quá lâu hoặc không thể tiếp tục di chuyển. Sự kiện này thường phản ánh tình trạng ùn tắc nghiêm trọng hoặc bế tắc giao thông trong mô phỏng. Vì vậy, đề án thiết kế hàm phần thưởng như sau:

$$r_i = -\left(|p_i| + 0.25 \cdot q_i^{in} + 0.01 \cdot \Delta w_i\right) - r^{dc}. \tag{2.8}$$

Trong đó:

$$r^{dc} = \frac{5.0 \cdot \Delta T_{dc}}{N_{policy}}. \tag{2.9}$$

Các thành phần của hàm phần thưởng có ý nghĩa như sau:

$|p_i|$: Giá trị tuyệt đối của áp lực cục bộ, đóng vai trò chính trong việc cân bằng lưu lượng vào và ra tại nút giao.

$0.25 \cdot q_i^{in}$: Thành phần phạt đối với hàng đợi đang dừng, nhằm giảm mức độ ùn tắc cục bộ.

$0.01 \cdot \Delta w_i$: Thành phần phạt đối với thời gian chờ tăng thêm, nhằm hạn chế sự gia tăng thời gian chờ tích lũy.

r^{dc} : Thành phần phạt do các sự kiện dịch chuyển bắt buộc, được phân bổ đều cho các tác tử điều khiển.

ΔT_{dc} : Số sự kiện dịch chuyển bắt buộc phát sinh thêm trong bước mô phỏng hiện tại.

N_{policy} : Tổng số tác tử điều khiển trong môi trường.

Hàm phần thưởng luôn nhận giá trị không dương. Khi trạng thái giao thông càng bất lợi, giá trị phần thưởng càng nhỏ. Cách thiết kế này biến bài toán học thành quá trình giảm dần các đại lượng bất lợi của hệ thống, đồng thời giúp việc diễn giải từng thành phần trong hàm phần thưởng trở nên trực quan hơn.

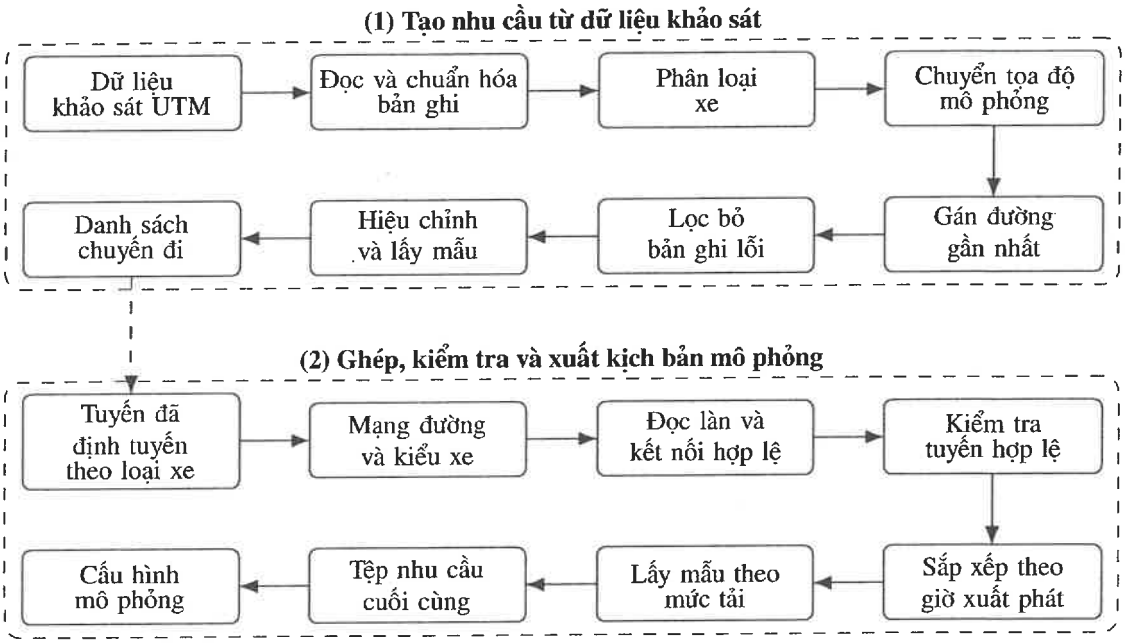
2.4 Môi trường mô phỏng

2.4.1 Mạng lưới và kịch bản

Mạng lưới giao thông trong đề án được xây dựng từ dữ liệu OSM của khu vực Thanh Xuân, Hà Nội và được xử lý bằng công cụ `netconvert` của SUMO. Kịch bản thực nghiệm được sử dụng là Thanh Xuân UTM, với mức nhu cầu mục tiêu khoảng 5.000 phương tiện mỗi giờ. Trong cấu hình mô phỏng, đề án kích hoạt mô hình chuyển làn phụ của SUMO dành cho xe máy nhằm phản ánh tốt hơn khả năng di chuyển linh hoạt của phương tiện hai bánh trong điều kiện giao thông hỗn hợp.

2.4.2 *Sinh nhu cầu giao thông từ dữ liệu UTM*

Nhu cầu giao thông được xây dựng từ dữ liệu khảo sát thuộc dự án UTM-Hanoi. Quy trình chuyển đổi từ dữ liệu khảo sát sang kịch bản mô phỏng SUMO được thực hiện theo các bước trình bày trong lưu đồ tại Hình 2.3.



Hình 2.3: Lưu đồ chuẩn bị dữ liệu đầu vào cho kịch bản thực nghiệm

Thành phần phương tiện được suy ra từ dữ liệu khảo sát phản ánh đặc trưng giao thông hỗn hợp của Hà Nội. Trong kịch bản Thanh Xuân UTM, mức nhu cầu mục tiêu là 5.000 phương tiện/giờ, gồm 3.663 xe máy, 1.243 ô tô và 94 xe buýt. Cần lưu ý rằng xe buýt trong kịch bản này được tổng hợp từ dữ liệu khảo sát và không đại diện cho lịch trình vận hành thực tế của các tuyến xe buýt.

2.4.3 *Hệ số PCU và loại phương tiện*

Để phản ánh tải trọng của dòng giao thông hỗn hợp, đề án sử dụng PCU nhằm quy đổi các loại phương tiện về cùng một đơn vị đo. Hệ số quy đổi được xác định dựa trên kiểu phương tiện trong SUMO. Cụ thể, xe máy, tương ứng với các kiểu phương tiện có chứa *motor*, được quy đổi bằng 0.5 PCU. Xe buýt, tương ứng với các kiểu phương tiện có chứa *bus*, được quy đổi bằng 3.0 PCU. Các phương tiện ô tô con còn lại được xem là nhóm mặc định và được quy đổi bằng 1.0 PCU.

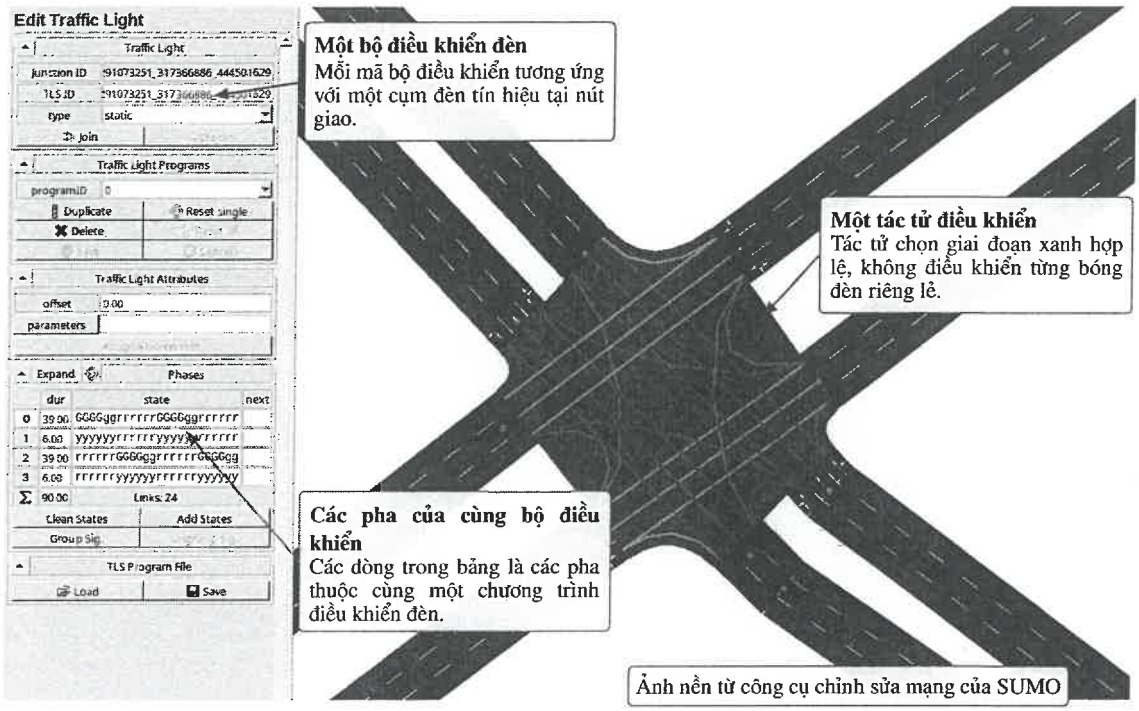
2.4.4 Tác tử điều khiển và bộ điều khiển cố định

Mỗi bộ điều khiển đèn tín hiệu trong mạng lưới được phân tích từ chương trình pha `tlLogic`. Một pha được xem là giai đoạn xanh khi pha đó chứa ít nhất một tín hiệu xanh, không xuất hiện tín hiệu vàng và không thuộc pha đỏ toàn bộ. Dựa trên tập các giai đoạn xanh, môi trường phân loại các bộ điều khiển đèn tín hiệu thành hai nhóm.

Nhóm thứ nhất là các tác tử điều khiển, gồm những bộ điều khiển có nhiều hơn một giai đoạn xanh. Đây là các nút giao có khả năng đưa ra quyết định điều khiển và được đưa vào quá trình học tăng cường. Nhóm thứ hai là các bộ điều khiển cố định, gồm những bộ điều khiển có không quá một giai đoạn xanh. Nhóm này vẫn tham gia mô phỏng trong SUMO, nhưng không nhận hành động từ chính sách học tăng cường.

Đối với mỗi tác tử điều khiển, số lượng hành động hợp lệ được xác định bằng $1+n_i$, trong đó n_i là số giai đoạn xanh của bộ điều khiển thứ i. Hành động đầu tiên tương ứng với việc duy trì giai đoạn hiện tại, trong khi các hành động còn lại đại diện cho yêu cầu chuyển sang một giai đoạn xanh khác. Tuy nhiên, trong quá trình thực thi, bộ lọc hành động chỉ cho phép tác tử chuyển sang giai đoạn kế tiếp theo thứ tự chu kỳ và chỉ khi điều kiện thời gian xanh tối thiểu đã được thỏa mãn.

Các bộ điều khiển cố định không được đưa vào tập tác tử học, không xuất hiện trong ma trận kề của đồ thị chính sách và cũng không tham gia vào quá trình ước lượng giá trị tập trung. Trong kịch bản Thanh Xuân UTM, môi trường cuối cùng sử dụng 41 tác tử điều khiển cho quá trình huấn luyện và đánh giá. Các tác tử này được kết nối theo quan hệ lân cận của mạng lưới giao thông; trong các biến thể GAT, thông tin giữa các nút giao lân cận được trao đổi gián tiếp thông qua bộ mã hóa đồ thị.



Hình 2.4: Minh họa một bộ điều khiển đèn tín hiệu trong SUMO/netedit và các pha thuộc cùng một tác tử điều khiển

2.5 Đồ thị chính sách

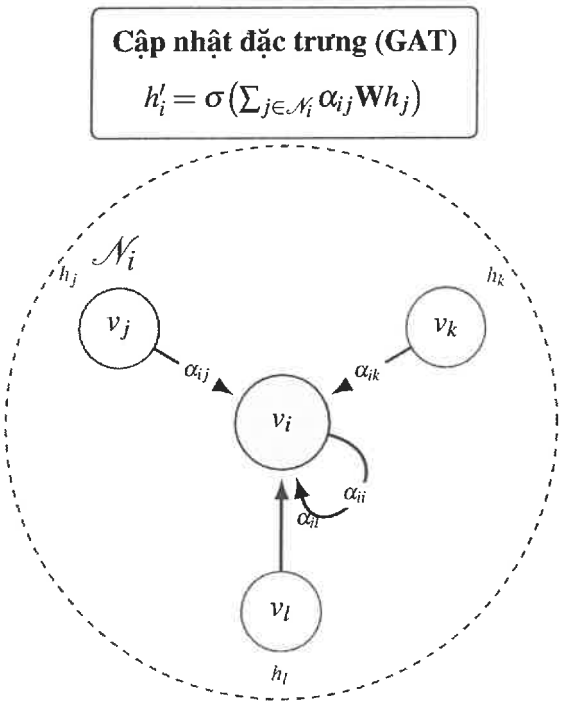
2.5.1 Xây dựng đồ thị

Với các biến thể sử dụng GAT, môi trường xây dựng một đồ thị chính sách $G = (V, E)$ từ mô tả mạng lưới giao thông trong SUMO. Đồ thị này chỉ bao gồm các tác tử điều khiển và được sử dụng để giới hạn phạm vi truyền tin giữa các nút trong bộ mã hóa đồ thị. Trong đồ thị chính sách, tập nút V biểu diễn tập các tác tử điều khiển. Mỗi nút $v_i \in V$ tương ứng với một bộ điều khiển đèn tín hiệu có nhiều hơn một giai đoạn xanh. Các bộ điều khiển cố định không được đưa vào đồ thị này. Tập cạnh E biểu diễn quan hệ lân cận giữa các tác tử điều khiển, trong đó một cạnh giữa v_i và v_j cho biết hai tác tử tương ứng có quan hệ kết nối trực tiếp trong mạng SUMO.

Cụ thể, hai tác tử được xem là lân cận nếu tồn tại ít nhất một đoạn đường kết nối trực tiếp giữa hai nút giao tương ứng trong mạng SUMO. Mặc dù các đoạn đường trong SUMO có hướng, ma trận kề dùng trong mô hình được đối xứng hóa để cho phép truyền tin hai chiều giữa các tác tử lân cận. Ngoài ra, ma trận kề $A \in 0, 1^{n \times n}$ được thêm cạnh tự nối để mỗi nút luôn giữ lại thông tin của chính nó trong quá trình

truyền tin:

$$A_{ij} = \begin{cases} 1, & \text{nếu } v_i \text{ và } v_j \text{ là hai tác tử lân cận, hoặc } i = j, \\ 0, & \text{trong các trường hợp còn lại.} \end{cases} \tag{2.10}$$



Hình 2.5: Cơ chế truyền tin trong GAT với trọng số chú ý động giữa các nút giao trong đồ thị chính sách

2.5.2 Cơ chế chú ý GAT

Bộ mã hóa đồ thị trong đề án gồm hai lớp GAT [21]. Tại mỗi lớp, đặc trưng của nút h_i trước hết được chiếu tuyến tính. Sau đó, trọng số chú ý giữa nút i và nút lân cận j được tính từ đặc trưng của hai nút như sau:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}_{src}^{\top} \mathbf{W} h_i + \mathbf{a}_{dst}^{\top} \mathbf{W} h_j\right) \tag{2.11}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}, \quad j \in \mathcal{N}_i \tag{2.12}$$

$$h_i' = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} h_j\right) \tag{2.13}$$

Trong các công thức trên, $\mathcal{N}_i = \{j \mid A_{ij} = 1\}$ là tập các nút được phép truyền tin tới nút i. Các cặp nút không lân cận bị loại khỏi bước chuẩn hóa trọng số chú ý nên không tham gia vào quá trình tổng hợp thông tin. Lớp GAT đầu tiên sử dụng H head song song và nối các đầu ra của các head này. Lớp GAT thứ hai sử dụng một head để đưa biểu diễn nút cuối cùng về chiều d_{gat}.

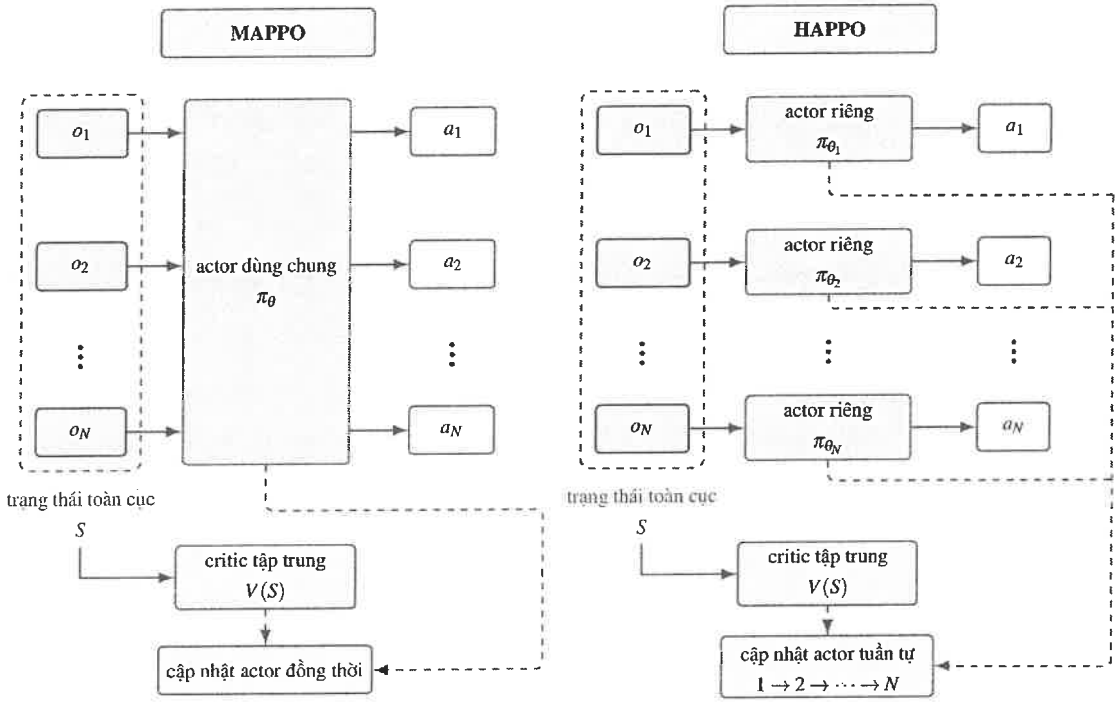
2.6 Kiến trúc các thuật toán

Trước khi trình bày từng thuật toán, cần làm rõ hai thành phần chính trong kiến trúc actor–critic. Actor là mạng chính sách, có nhiệm vụ nhận quan sát của tác tử và sinh phân phối xác suất trên các hành động hợp lệ. Từ phân phối này, tác tử lựa chọn hành động điều khiển tại mỗi bước thời gian. Critic là mạng giá trị, có nhiệm vụ ước lượng mức độ tốt của trạng thái và cung cấp tín hiệu hỗ trợ cho quá trình cập nhật actor.

Trong các thuật toán được triển khai, actor có thể được tổ chức theo hai cách. Với MAPPO, các tác tử sử dụng chung một actor; với HAPPO, mỗi tác tử có một actor riêng. Ở cả hai trường hợp, critic đều được thiết kế theo hướng tập trung, tức là có thể khai thác thông tin toàn cục trong giai đoạn huấn luyện để hỗ trợ ước lượng giá trị.

Các thuật toán cũng khác nhau ở cách cập nhật actor. Trong cơ chế cập nhật đồng thời, dữ liệu của tất cả tác tử được gom vào cùng một batch và được sử dụng để tối ưu hóa chính sách trong cùng một lượt cập nhật. Trong cơ chế cập nhật tuần tự, các actor riêng được cập nhật lần lượt theo một thứ tự cố định; khi cập nhật một actor, thuật toán có xét đến ảnh hưởng của các actor đã được cập nhật trước đó.

Hình 2.6 minh họa khác biệt chính giữa MAPPO và HAPPO. MAPPO sử dụng một actor chung và cập nhật đồng thời, trong khi HAPPO sử dụng actor riêng cho từng tác tử và cập nhật tuần tự.



Hình 2.6: So sánh cơ chế actor và cập nhật của MAPPO và HAPPO

2.6.1 MAPPO

MAPPO [17] là biến thể của PPO trong môi trường đa tác tử. Trong đề án này, MAPPO sử dụng một mạng chính sách chung cho tất cả các tác tử điều khiển. Mạng chính sách này nhận quan sát cục bộ của từng nút giao và sinh phân phối xác suất trên các hành động hợp lệ của nút đó. Mạng giá trị được thiết kế theo hướng tập trung: thay vì chỉ nhận quan sát của một tác tử riêng lẻ, critic sử dụng trạng thái toàn cục được tạo bằng cách vector hóa đặc trưng của toàn bộ tác tử để ước lượng giá trị trong quá trình huấn luyện.

2.6.1.1 Kiến trúc

Trong cấu hình mặc định của đề án, actor của MAPPO nhận trực tiếp vector quan sát cục bộ $o_i \in \mathbb{R}^8$ của từng tác tử và đưa qua hai lớp MLP ẩn có kích thước 128. Tầng cuối sinh ra các điểm số hành động chưa chuẩn hóa tương ứng với không gian hành động đã được đưa về cùng một kích thước cực đại. Việc chuẩn hóa kích thước này là cần thiết vì các nút giao có thể có số lượng giai đoạn xanh khác nhau. Trước bước lấy mẫu, bộ lọc hành động loại bỏ các hành động không hợp lệ, nhờ đó actor chỉ lựa chọn và học trên tập hành động hợp lệ tại từng thời điểm.

Critic của MAPPO là mạng giá trị tập trung. Đầu vào của critic là trạng thái toàn

cục được tạo bằng cách vector hóa toàn bộ đặc trưng của các tác tử điều khiển. Với 41 tác tử và vector quan sát 8 chiều, critic xử lý một vector có kích thước $41 \times 8 = 328$. Vector này được đưa qua MLP ẩn có kích thước 256 để ước lượng giá trị dùng trong quá trình huấn luyện.

2.6.1.2 Cập nhật tham số

MAPPO tối ưu hóa chính sách của tất cả tác tử thông qua một bộ trọng số actor chung. Sau mỗi lượt thu thập mẫu, hệ thống ghi nhận quan sát, hành động, log-xác suất cũ, phần thưởng và trạng thái toàn cục. Dựa trên giá trị do critic tập trung ước lượng, hệ thống tính GAE cho dữ liệu tương tác theo thời gian của từng tác tử, sau đó chuẩn hóa lợi thế trên toàn bộ lô dữ liệu trước khi thực hiện nhiều vòng cập nhật theo mục tiêu PPO-Clip.

Chia sẻ tham số là một đặc điểm quan trọng của MAPPO trong đề án này. Thiết kế này giúp giảm số lượng tham số của actor và cho phép kinh nghiệm từ tất cả các nút giao cùng đóng góp vào quá trình cập nhật một mạng chính sách chung. Nhờ đó, mô hình có thể học các quy luật điều khiển chung của dòng giao thông hỗn hợp từ dữ liệu của nhiều nút giao. Tuy nhiên, cách tiếp cận này cũng giả định rằng các tác tử có thể sử dụng cùng một họ chính sách, nên khả năng chuyên biệt hóa theo từng nút giao bị giới hạn.

2.6.2 HAPPO

HAPPO [18] là biến thể mở rộng của MAPPO theo hướng cho phép mỗi tác tử điều khiển sở hữu một mạng chính sách riêng. Thay vì sử dụng một actor chung cho tất cả các nút giao như MAPPO, HAPPO gán cho mỗi tác tử một actor MLP độc lập và cập nhật các actor này theo thứ tự tuần tự. Mạng giá trị vẫn được thiết kế dưới dạng critic tập trung dùng chung, nhận trạng thái toàn cục để hỗ trợ quá trình ước lượng lợi thế trong giai đoạn huấn luyện.

2.6.2.1 Kiến trúc

Khác với MAPPO, HAPPO không chia sẻ trọng số actor giữa các tác tử. Mỗi tác tử điều khiển có một actor MLP riêng, nhận đầu vào là quan sát cục bộ của chính tác tử đó và sinh phân phối xác suất trên các hành động hợp lệ. Thiết kế này cho phép các nút giao có cấu trúc pha, vai trò trong mạng lưới và mẫu lưu lượng khác nhau học được các chính sách chuyên biệt hơn.

Mặc dù các actor được tách riêng, HAPPO vẫn sử dụng một critic tập trung trên trạng thái toàn cục. Cách tổ chức này giữ lại lợi ích của khung CTDE: actor chỉ cần

quan sát cục bộ khi thực thi, trong khi critic có thể sử dụng thông tin toàn cục trong quá trình huấn luyện để tạo tín hiệu học ổn định hơn.

2.6.2.2 *Cập nhật tuần tự*

Trọng tâm của HAPPO là cơ chế cập nhật tuần tự. Trong mỗi epoch, hệ thống không cập nhật đồng thời tất cả các actor mà lần lượt cập nhật từng actor theo một thứ tự cố định của danh sách tác tử điều khiển. Khi actor của tác tử thứ i được cập nhật, các tác tử trước đó đã sử dụng chính sách mới, trong khi các tác tử phía sau vẫn giữ chính sách cũ.

Để ghi nhận ảnh hưởng của các tác tử đã được cập nhật trước đó, HAPPO bổ sung tỷ lệ tích lũy M_i vào mục tiêu tối ưu của tác tử hiện tại:

$$L_i = \mathbb{E}\left[\min\left(M_i \cdot \rho_i(\theta)\hat{A}_i,\ \text{clip}(M_i \cdot \rho_i(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_i\right)\right]. \tag{2.14}$$

Trong đó, M_i được tính từ tỷ lệ xác suất của các tác tử đã cập nhật trước đó:

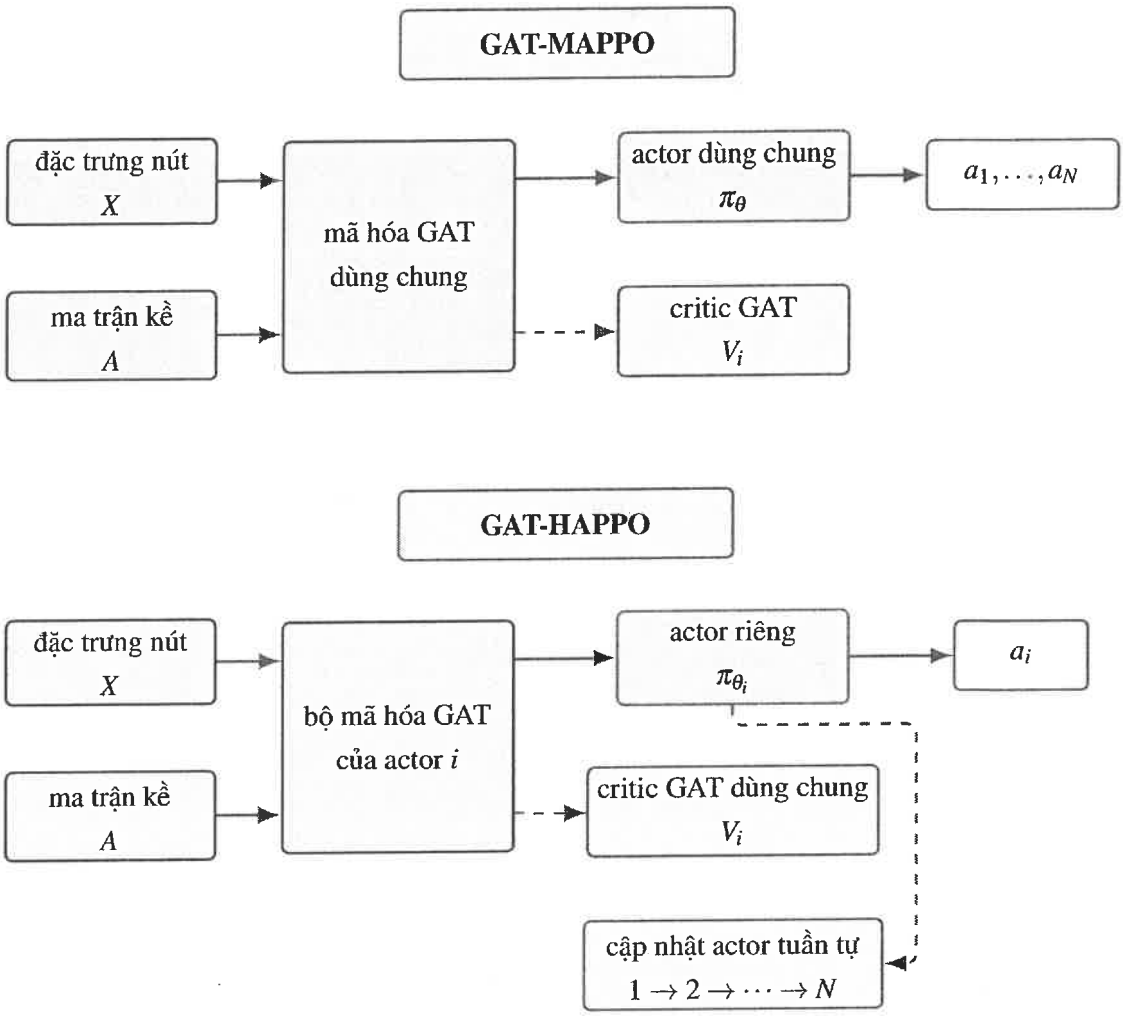
$$M_i = \prod_{k=1}^{i-1} \rho_k(\theta). \tag{2.15}$$

Sau khi cập nhật xong mạng chính sách của tác tử i, tỷ lệ tích lũy cho tác tử kế tiếp được tính lại dựa trên log-xác suất dưới chính sách mới. Về trực giác, cơ chế này giúp tác tử hiện tại được tối ưu trong bối cảnh đã xét đến những thay đổi chính sách của các tác tử được cập nhật trước đó, thay vì giả định rằng chính sách của toàn bộ hệ thống vẫn giữ nguyên.

Cách cập nhật này giúp giảm tính không dừng do nhiều tác tử cùng thay đổi chính sách quá mạnh trong một bước tối ưu hóa. Bù lại, chi phí huấn luyện tăng lên vì mỗi epoch phải lần lượt cập nhật nhiều actor riêng biệt. Do đó, HAPPO phù hợp hơn trong các bài toán có tác tử không đồng nhất, nơi từng nút giao cần học hành vi điều khiển riêng, nhưng phải đánh đổi bằng thời gian huấn luyện và bộ nhớ mô hình lớn hơn.

2.6.3 GAT-MAPPO

Hình 2.7 minh họa hai biến thể thuật toán sử dụng GAT. Cả GAT-MAPPO và GAT-HAPPO đều nhận ma trận đặc trưng nút và ma trận kề làm đầu vào để khai thác cấu trúc liên kết giữa các nút giao. Tuy nhiên, hai biến thể này khác nhau ở cách tổ chức actor: GAT-MAPPO sử dụng actor chung cho tất cả tác tử, trong khi GAT-HAPPO sử dụng actor riêng cho từng tác tử và cập nhật các actor theo cơ chế tuần tự.



Hình 2.7: So sánh kiến trúc và cơ chế truyền tin đồ thị trong GAT-MAPPO và GAT-HAPPO

GAT-MAPPO giữ nguyên cơ chế cập nhật của MAPPO, tức là sử dụng mạng chính sách dùng chung và cập nhật đồng thời các tác tử. Điểm khác biệt chính nằm ở bộ mã hóa đầu vào. Thay vì xử lý từng vector quan sát cục bộ bằng MLP, GAT-MAPPO sử dụng bộ mã hóa đồ thị GAT cho cả mạng chính sách và mạng giá trị. Nhờ đó, quyết định của mỗi tác tử không chỉ dựa trên quan sát cục bộ của chính nó mà còn có thể khai thác thông tin từ các nút giao lân cận trong mạng lưới.

Cụ thể, cả mạng chính sách và mạng giá trị đều nhận hai đầu vào là ma trận đặc trưng nút X và ma trận kề A của đồ thị chính sách. Ma trận X chứa đặc trưng quan sát của các tác tử điều khiển, còn ma trận A mô tả quan hệ lân cận giữa các nút giao. Trong mạng chính sách, bộ mã hóa GAT biến đổi ma trận đặc trưng có kích thước

$(N \times d_{in})$ thành ma trận biểu diễn nút có kích thước $(N \times d_{gat})$. Từ biểu diễn này, mỗi tác tử sử dụng hàng tương ứng với nó để sinh điểm số hành động. Trong mạng giá trị, bộ mã hóa GAT xử lý toàn bộ đặc trưng nút trên đồ thị và xuất giá trị V_i cho từng tác tử điều khiển.

Mục tiêu PPO-Clip được giữ nguyên như trong MAPPO. Critic của GAT-MAPPO xuất giá trị V_i cho từng tác tử điều khiển; từ các giá trị này, GAE được tính độc lập trên dữ liệu tương tác theo thời gian của từng tác tử.

2.6.3.1 Biểu diễn đầu vào và truyền tin

Trong biến thể này, đầu vào của actor không còn là từng vector quan sát riêng lẻ mà là toàn bộ ma trận đặc trưng nút $X \in \mathbb{R}^{N \times d_{in}}$. Với các biến thể GAT, $d_{in} = 9$ do vector quan sát được bổ sung thêm đặc trưng về tỷ lệ số pha xanh của tác tử so với tác tử có nhiều pha xanh nhất.

Bộ mã hóa GAT gồm hai lớp và thực hiện truyền tin giữa các nút lân cận theo ma trận kề A. Sau bước truyền tin, biểu diễn của tác tử i không chỉ phản ánh trạng thái tại nút giao của nó mà còn chứa thông tin tóm tắt từ các nút giao lân cận. Điều này giúp tác tử đưa ra quyết định trong bối cảnh mạng lưới rộng hơn, thay vì chỉ dựa trên quan sát cục bộ ban đầu.

Điểm khác biệt quan trọng so với MAPPO thuần MLP là actor và critic của GAT-MAPPO đều khai thác cấu trúc liên kết của mạng lưới giao thông. Thay vì làm phẳng toàn bộ trạng thái thành một vector dài cố định, critic đồ thị học trực tiếp trên cấu trúc đồ thị, phù hợp hơn với bài toán có quan hệ phụ thuộc không gian giữa các nút giao.

2.6.3.2 Ưu điểm và chi phí

GAT-MAPPO giữ được ưu điểm mở rộng của MAPPO nhờ sử dụng actor dùng chung, đồng thời bổ sung khả năng phối hợp cục bộ thông qua cơ chế truyền tin trên đồ thị. Ví dụ, khi một nút giao hạ lưu đang ùn ứ, thông tin này có thể được truyền tới nút thượng lưu lân cận thông qua GAT, giúp nút thượng lưu điều chỉnh xác suất chọn hành động phù hợp hơn.

Đổi lại, chi phí tính toán của GAT-MAPPO cao hơn MAPPO thuần MLP, vì mô hình phải tính trọng số chú ý trên các cạnh của đồ thị thay vì xử lý độc lập từng vector quan sát bằng MLP.

2.6.4 GAT-HAPPO

GAT-HAPPO kết hợp hai ý tưởng chính: cơ chế cập nhật tuần tự của HAPPO và khả năng truyền tin trên đồ thị của GAT. Tương tự HAPPO, mỗi tác tử điều khiển có

một mạng chính sách riêng. Tuy nhiên, thay vì chỉ xử lý quan sát cục bộ bằng MLP, mỗi actor sử dụng bộ mã hóa GAT để khai thác thêm thông tin từ các nút giao lân cận trong đồ thị chính sách. Mạng giá trị vẫn là mạng GAT tập trung dùng chung, tương tự như trong GAT-MAPPO.

2.6.4.1 Kết hợp chính sách riêng và biểu diễn đồ thị

Trong bốn thuật toán được triển khai, GAT-HAPPO là biến thể có khả năng biểu diễn cao nhất. Thuật toán này cho phép mỗi tác tử có một actor riêng như trong HAPPO, nhờ đó các nút giao khác nhau có thể học các chính sách điều khiển chuyên biệt. Đồng thời, actor của từng tác tử vẫn sử dụng biểu diễn nút được tạo bởi bộ mã hóa GAT, nên quyết định điều khiển có thể phản ánh thêm bối cảnh không gian từ các nút giao lân cận.

Cụ thể, bộ mã hóa GAT nhận ma trận đặc trưng nút X và ma trận kề A của đồ thị chính sách, sau đó tạo ra biểu diễn mới cho toàn bộ tác tử. Khi cập nhật hoặc ra quyết định cho tác tử i, mô hình chỉ sử dụng biểu diễn tương ứng với tác tử đó để sinh điểm số hành động. Nhờ vậy, mỗi tác tử vừa có chính sách chuyên biệt, vừa có khả năng phản ứng với trạng thái của vùng giao thông xung quanh.

Thiết kế này phù hợp với các mạng lưới có nút giao không đồng nhất, chẳng hạn khác nhau về số pha xanh, cấu trúc làn đường, vai trò trong mạng lưới hoặc mẫu lưu lượng. Trong trường hợp đó, một chính sách dùng chung có thể chưa đủ để biểu diễn đầy đủ sự khác biệt giữa các nút giao, trong khi chính sách riêng có kết hợp thông tin đồ thị có thể thích ứng tốt hơn với từng vị trí trong mạng lưới.

2.6.4.2 Đánh đổi thực thi

GAT-HAPPO cũng là biến thể có chi phí tính toán cao nhất trong bốn thuật toán. So với GAT-MAPPO, hệ thống phải duy trì nhiều actor riêng biệt và cập nhật các actor này theo thứ tự tuần tự. So với HAPPO, mỗi actor lại cần thêm bước mã hóa đồ thị bằng GAT. Vì vậy, thuật toán này có khả năng biểu diễn tốt hơn, nhưng đồng thời đòi hỏi nhiều tài nguyên tính toán hơn, thời gian huấn luyện dài hơn và nhạy hơn với lựa chọn siêu tham số.

Bảng 2.1 tóm tắt sự khác biệt giữa bốn thuật toán theo hai khía cạnh chính: dạng đầu vào được sử dụng và cách tổ chức mạng chính sách, mạng giá trị.

Bảng 2.1: So sánh kiến trúc bốn biến thể thuật toán được triển khai

Thuật toán	Đầu vào	Mạng chính sách	Mạng giá trị
MAPPO	$\{o_i\}$	MLP dùng chung, đầu ra riêng theo tác tử	MLP tập trung trên trạng thái toàn cục đã vector hóa
HAPPO	$\{o_i\}$	MLP riêng cho từng tác tử, cập nhật tuần tự	MLP tập trung dùng chung
GAT-MAPPO	X, A	GAT dùng chung, truyền tin qua đồ thị chính sách	GAT tập trung, xuất V_i cho từng tác tử
GAT-HAPPO	X, A	GAT riêng cho từng tác tử, cập nhật tuần tự	GAT tập trung dùng chung

Trong Bảng 2.1, X là ma trận đặc trưng của các tác tử điều khiển và A là ma trận kề của đồ thị chính sách.

2.7 Thiết lập thông số huấn luyện

Bảng 2.2 liệt kê các thông số chính được sử dụng trong quá trình huấn luyện. Các giá trị này được giữ cố định giữa các thuật toán nhằm bảo đảm điều kiện so sánh nhất quán. Đề án không tối ưu hóa siêu tham số riêng cho từng thuật toán; do đó, kết quả được diễn giải trong phạm vi cấu hình thực nghiệm đã chọn.

Bảng 2.2: Các thông số chính trong cấu hình huấn luyện

Nhóm thông số	**Giá trị sử dụng**
Bộ tối ưu	Adam
Tốc độ học	3×10^{-4}
Hệ số chiết khấu (γ)	0.99
Hệ số GAE (λ)	0.95
Ngưỡng cắt PPO (ε)	0.2
Hệ số entropy	0.01
Hệ số hàm giá trị	0.5
Chuẩn gradient tối đa	0.5
Kích thước ẩn của actor MLP	128
Kích thước ẩn của critic MLP	256
Kích thước biểu diễn GAT	64
Số đầu chú ý GAT ở lớp đầu	4
Số bước thu thập mẫu mỗi lượt	128
Số vòng cập nhật mỗi lượt	4
Kích thước lô nhỏ	32
Bước quyết định điều khiển	5 giây
Thời gian xanh tối thiểu	10 giây
Thời lượng một lần mô phỏng	3600 giây

2.8 Giao thức lưu mô hình và bảo đảm tính tái lập

Trong quá trình huấn luyện, mỗi mô hình được lưu kèm với các thông tin cấu hình đã tạo ra mô hình đó. Các thông tin này bao gồm mạng đường, dữ liệu nhu cầu giao thông, cấu hình mô phỏng, siêu tham số huấn luyện, hạt giống ngẫu nhiên, danh sách tác tử điều khiển và cấu trúc đồ thị chính sách. Việc lưu kèm cấu hình giúp mô hình không bị tách rời khỏi kịch bản thực nghiệm ban đầu.

Khi tiếp tục huấn luyện hoặc đánh giá một bản lưu mô hình, hệ thống so sánh cấu hình hiện tại với cấu hình được ghi nhận tại thời điểm lưu mô hình. Nếu hai cấu hình không khớp, chẳng hạn khác mạng đường, khác dữ liệu nhu cầu giao thông hoặc khác danh sách tác tử điều khiển, mô hình sẽ không được sử dụng cho kịch bản hiện tại. Cơ chế này giúp tránh trường hợp đánh giá nhầm một bản lưu mô hình trên một môi trường không tương thích, từ đó hạn chế sai lệch trong kết quả thực nghiệm.

Cách tổ chức này góp phần tăng tính tái lập của thí nghiệm. Khi mô hình, cấu hình mô phỏng và hạt giống ngẫu nhiên được lưu giữ nhất quán, các lần chạy sau có thể kiểm tra lại kết quả trên cùng một điều kiện thực nghiệm. Đồng thời, bản ghi cấu hình cũng giúp truy vết nguồn gốc của từng bản lưu mô hình khi so sánh kết quả giữa các thuật toán.

2.9 Minh họa cấu hình chương trình đèn tín hiệu trong SUMO

Để làm rõ cách một bộ điều khiển đèn tín hiệu được mô tả trong SUMO, đoạn mã dưới đây minh họa cấu trúc cơ bản của một chương trình đèn tín hiệu trong tệp XML. Mỗi chương trình đèn tín hiệu có một định danh, một kiểu điều khiển, một mã chương trình và một danh sách các pha. Trước khi được tác tử điều khiển trong môi trường học tăng cường, các pha này đóng vai trò là cấu hình tín hiệu mặc định của nút giao.

Mã nguồn 2.1: Cấu hình logic đèn tín hiệu

```
<tlLogic id="J1" type="static" programID="0" offset="0">
    <phase duration="33" state="GGgrrrGGgrrr"/>
    <phase duration="3"  state="yyyrrryyyrrr"/>
    <phase duration="33" state="rrrGGgrrrGGg"/>
    <phase duration="3"  state="rrryyyrrryyy"/>
</tlLogic>
```

2.10 Kết luận chương

Chương này đã trình bày mô hình học tăng cường đa tác tử cho bài toán điều khiển đèn tín hiệu. Nội dung chương bao gồm nền tảng học tăng cường đa tác tử, cơ chế cập nhật chính sách bằng PPO và GAE, cách mô hình hóa bài toán dưới dạng trò chơi Markov, quy trình xây dựng môi trường SUMO từ dữ liệu khảo sát UTM, đồ thị chính sách và kiến trúc của bốn thuật toán được triển khai. Các nội dung này tạo cơ sở cho phần thực nghiệm ở chương tiếp theo, trong đó các phương pháp điều khiển được đánh giá trên nhiều hạt giống ngẫu nhiên trong kịch bản Thanh Xuân UTM.

Chương 3
THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Chương này trình bày thiết lập thực nghiệm và kết quả đánh giá các phương pháp điều khiển đèn tín hiệu trên kịch bản Thanh Xuân UTM. Mục tiêu của chương là đánh giá các thuật toán học tăng cường đa tác tử được trình bày tại Mục 2.6 so với hai phương pháp đối chứng: điều khiển cố định và Max-Pressure. Các kết quả được diễn giải theo từng chỉ số vận hành, thay vì giả định tồn tại một phương pháp vượt trội trên toàn bộ tiêu chí.

3.1 Kịch bản và giao thức đánh giá

3.1.1 Kịch bản Thanh Xuân UTM

Kịch bản thực nghiệm chính mô phỏng mạng lưới giao thông khu vực Thanh Xuân, Hà Nội trong SUMO. Mạng đường được xây dựng từ dữ liệu OSM, trong khi nhu cầu giao thông được sinh từ dữ liệu khảo sát UTM-Hanoi và điều chỉnh về mức nhu cầu mục tiêu 5.000 phương tiện/giờ. Cấu hình mô phỏng sử dụng mô hình chuyển làn phụ nhằm phản ánh tốt hơn đặc trưng giao thông hỗn hợp có tỷ lệ xe máy cao.

Bảng 3.1: Cấu hình kịch bản thực nghiệm Thanh Xuân UTM

Tham số	**Giá trị**
Khu vực mô phỏng	Thanh Xuân, Hà Nội
Nguồn mạng lưới	OSM, xử lý bằng công cụ `netconvert` của SUMO
Nguồn nhu cầu	Dữ liệu khảo sát UTM-Hanoi
Mức nhu cầu mục tiêu	5.000 phương tiện/giờ
Thời lượng mỗi lần đánh giá	3.600 giây mô phỏng
Bước quyết định	5 giây
Số hạt giống đánh giá	10
Tác tử điều khiển	41 tác tử
Thành phần phương tiện	Xe máy, ô tô con và xe buýt, suy ra từ dữ liệu khảo sát
Mô hình mô phỏng	Mô phỏng vi mô trong SUMO, có bật mô hình chuyển làn phụ cho xe máy



Hình 3.1: Mạng lưới giao thông khu vực Thanh Xuân trong môi trường SUMO

Quy trình chuẩn bị mạng lưới, sinh nhu cầu giao thông và xây dựng kịch bản mô phỏng được trình bày tại Hình 2.3. Trong chương này, các kết quả đánh giá được lấy từ cùng một giao thức thực nghiệm, cùng kịch bản, cùng mức nhu cầu và cùng tập hạt giống ngẫu nhiên để bảo đảm khả năng so sánh giữa các phương pháp.

3.1.2 Các phương pháp so sánh

Trong kịch bản thực nghiệm, đề án đánh giá sáu phương pháp điều khiển trong cùng điều kiện mô phỏng. Phương pháp điều khiển cố định sử dụng chương trình đèn tín hiệu mặc định và không có cơ chế thích ứng theo trạng thái giao thông hiện thời. Max-Pressure là phương pháp điều khiển thích ứng dựa trên nguyên lý áp lực cực đại, trong đó bộ điều khiển lựa chọn giai đoạn tín hiệu có áp lực lớn nhất tại mỗi bước quyết định.

Bên cạnh hai phương pháp đối chứng, đề án đánh giá bốn thuật toán học tăng cường đa tác tử đã được trình bày ở Chương 2. MAPPO sử dụng actor MLP dùng chung cho các tác tử và critic MLP tập trung. HAPPO sử dụng actor MLP riêng cho từng tác tử và cập nhật chính sách theo thứ tự tuần tự. GAT-MAPPO mở rộng MAPPO bằng cách sử dụng bộ mã hóa GAT để khai thác quan hệ lân cận trong đồ thị chính sách.

GAT-HAPPO kết hợp actor riêng của HAPPO với biểu diễn đồ thị từ GAT.

Trong các phương pháp so sánh, Max-Pressure được xem là phương pháp đối chứng thích ứng quan trọng. Vì vậy, kết quả của các thuật toán học tăng cường không chỉ được so sánh với điều khiển cố định, mà còn được đối chiếu trực tiếp với Max-Pressure để đánh giá mức cải thiện trong một bối cảnh đối chứng mạnh hơn.

3.1.3 Giao thức huấn luyện và đánh giá

Các thuật toán học tăng cường được huấn luyện trên cùng kịch bản Thanh Xuân UTM với mức nhu cầu 5.000 phương tiện/giờ. Mỗi lần mô phỏng huấn luyện kéo dài 3.600 giây. Do tác tử đưa ra quyết định sau mỗi 5 giây mô phỏng, mỗi lần mô phỏng huấn luyện gồm 720 bước quyết định. Ở mỗi bước quyết định, môi trường thu thập trạng thái giao thông từ SUMO, tạo quan sát cho 41 tác tử điều khiển, áp dụng bộ lọc hành động hợp lệ, thực thi hành động điều khiển và tính phần thưởng tương ứng. Dữ liệu tương tác theo thời gian sau đó được sử dụng để tính lợi thế bằng GAE và cập nhật chính sách theo PPO.

Trong thực nghiệm chính, mỗi thuật toán học tăng cường được huấn luyện trong 500 lần mô phỏng, tương ứng với 360.000 bước quyết định cho mỗi mô hình. Sau quá trình huấn luyện, bản lưu mô hình cuối cùng của mỗi thuật toán được sử dụng cho giai đoạn đánh giá. Mỗi bản lưu mô hình chứa trọng số mô hình cùng với thông tin cấu hình thực nghiệm, bao gồm kịch bản mô phỏng, dữ liệu nhu cầu, danh sách tác tử điều khiển, siêu tham số huấn luyện và cấu trúc đồ thị chính sách. Việc lưu kèm cấu hình giúp kiểm tra sự tương thích giữa mô hình đã huấn luyện và kịch bản được sử dụng khi đánh giá.

Các lần chạy thực nghiệm được thực hiện trên máy tính cá nhân sử dụng CPU Intel Core i5-12400 và 32 GB RAM. Thời gian huấn luyện và tần suất lưu các bản lưu mô hình trung gian không được ghi nhận đầy đủ trong bản ghi thực nghiệm. Vì vậy, đề án không sử dụng các bản lưu mô hình trung gian để lựa chọn mô hình tốt nhất, mà chỉ sử dụng bản lưu mô hình cuối cùng của mỗi thuật toán cho giai đoạn đánh giá.

Giai đoạn đánh giá được tách khỏi giai đoạn huấn luyện. Trong giai đoạn này, mô hình được nạp từ bản lưu mô hình đã lưu, chính sách chỉ được dùng để chọn hành động và không cập nhật trọng số. Mỗi phương pháp được đánh giá trên 10 hạt giống ngẫu nhiên với cùng thời lượng mô phỏng 3.600 giây. Hai phương pháp đối chứng không học, gồm điều khiển cố định và Max-Pressure, cũng được chạy trên cùng tập hạt giống để bảo đảm điều kiện so sánh nhất quán. Kết quả cuối cùng được tổng hợp

bằng giá trị trung bình và độ lệch chuẩn trên 10 lần chạy.

3.2 Chỉ số đánh giá

Đề án sử dụng bốn chỉ số chính để đánh giá hiệu quả điều khiển. Chỉ số thứ nhất là số phương tiện hoàn thành hành trình trong 3.600 giây mô phỏng. Giá trị này càng cao thì năng lực giải tỏa của mạng lưới càng tốt. Chỉ số thứ hai là thời gian chờ trung bình của các chuyến đi đã hoàn thành. Giá trị này càng thấp thì chất lượng di chuyển của phương tiện càng được cải thiện. Chỉ số thứ ba là tốc độ trung bình của phương tiện trong mô phỏng. Giá trị tốc độ cao hơn thường phản ánh trạng thái vận hành thông thoáng hơn. Chỉ số thứ tư là số sự kiện dịch chuyển bắt buộc, tức số lần SUMO buộc phải dịch chuyển phương tiện khỏi trạng thái bế tắc trong mô phỏng.

Bốn chỉ số trên phản ánh các khía cạnh khác nhau của hệ thống giao thông. Số phương tiện hoàn thành hành trình phản ánh năng lực phục vụ của mạng lưới, trong khi thời gian chờ và tốc độ trung bình gắn trực tiếp hơn với chất lượng di chuyển. Số sự kiện dịch chuyển bắt buộc được sử dụng như một chỉ số phụ để theo dõi mức độ bế tắc trong mô phỏng. Chỉ số này không được dùng làm bằng chứng chính về hiệu quả điều khiển, vì nó có thể chịu ảnh hưởng của hình học mạng đường, cấu hình mô phỏng và dao động giữa các hạt giống ngẫu nhiên.

3.3 Kết quả đánh giá đa hạt giống

3.3.1 Bảng kết quả chính

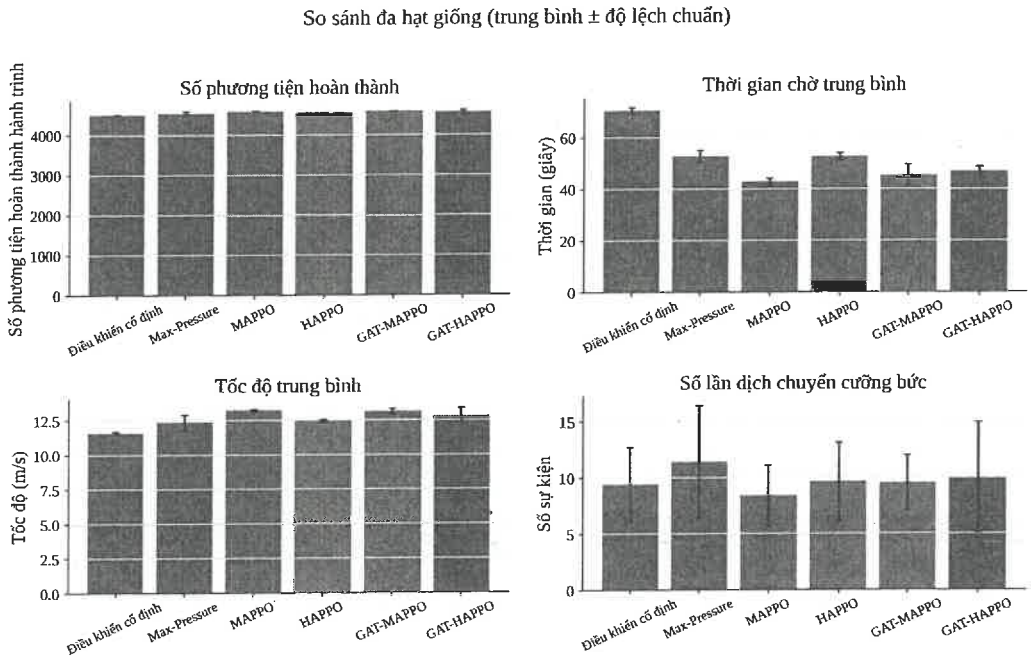
Bảng 3.2 trình bày kết quả trung bình và độ lệch chuẩn trên 10 hạt giống đánh giá. Tất cả phương pháp được chạy trên cùng kịch bản Thanh Xuân UTM, cùng mức nhu cầu 5.000 phương tiện/giờ và cùng thời lượng mô phỏng 3.600 giây.

Bảng 3.2: Kết quả trung bình ± độ lệch chuẩn trên 10 hạt giống đánh giá, kịch bản Thanh Xuân UTM

Phương pháp	**Số phương tiện hoàn thành (xe)**	**Chờ TB (s)**	**Tốc độ TB (m/s)**	**Dịch chuyển bắt buộc**
Điều khiển cố định	4495.3 ± 8.9	70.47 ± 0.90	11.55 ± 0.09	8.1 ± 3.4
Max-Pressure	4543.0 ± 35.5	52.03 ± 1.80	12.48 ± 0.40	9.9 ± 4.5
MAPPO	4580.5 ± 7.1	42.79 ± 1.36	13.19 ± 0.09	8.4 ± 2.7
HAPPO	4544.1 ± 8.1	52.49 ± 1.50	12.44 ± 0.10	9.6 ± 3.5
GAT-MAPPO	4583.3 ± 6.0	45.26 ± 4.09	13.10 ± 0.21	9.5 ± 2.5
GAT-HAPPO	4562.3 ± 48.9	46.88 ± 1.92	12.79 ± 0.57	10.0 ± 4.9

Kết quả cho thấy các phương pháp học tăng cường, đặc biệt là MAPPO và GAT-MAPPO, cho kết quả tốt hơn rõ ràng so với điều khiển cố định ở các chỉ số vận hành chính. So với điều khiển cố định, MAPPO giảm thời gian chờ trung bình từ 70.47 giây xuống 42.79 giây, tương đương giảm khoảng 39.3%. Đồng thời, tốc độ trung bình tăng từ 11.55 m/s lên 13.19 m/s, tương đương tăng khoảng 14.2%.

So với Max-Pressure, MAPPO có thêm 37.5 phương tiện hoàn thành hành trình, giảm thời gian chờ trung bình 9.24 giây và tăng tốc độ trung bình 0.72 m/s. Mức giảm thời gian chờ từ 52.03 giây xuống 42.79 giây tương đương khoảng 17.8%. GAT-MAPPO có số phương tiện hoàn thành hành trình cao nhất, với 4583.3 xe, nhưng chỉ nhỉnh hơn MAPPO 2.8 xe. Trong khi đó, thời gian chờ và tốc độ trung bình của GAT-MAPPO kém hơn MAPPO.



Hình 3.2: So sánh bốn chỉ số đánh giá chính trên kịch bản Thanh Xuân UTM

3.3.2 So sánh với Max-Pressure

Bảng 3.3 trình bày độ chênh lệch trung bình của từng phương pháp so với Max-Pressure. Dấu dương ở số phương tiện hoàn thành và tốc độ là tốt hơn, trong khi dấu âm ở thời gian chờ và số sự kiện dịch chuyển bắt buộc là tốt hơn.

Bảng 3.3: Độ chênh lệch trung bình so với Max-Pressure

Phương pháp	**Δ hoàn thành (xe)**	**Δ chờ (s)**	**Δ tốc độ (m/s)**	**Δ dịch chuyển bắt buộc**
Điều khiển cố định	-47.7	+18.44	-0.92	-1.8
MAPPO	+37.5	-9.24	+0.72	-1.5
HAPPO	+1.1	+0.46	-0.03	-0.3
GAT-MAPPO	+40.3	-6.77	+0.63	-0.4
GAT-HAPPO	+19.3	-5.15	+0.31	+0.1

MAPPO và GAT-MAPPO là hai phương pháp cho kết quả tốt hơn Max-Pressure ở ba chỉ số vận hành chính. Tuy nhiên, số phương tiện hoàn thành hành trình chỉ tăng khoảng 37 - 40 xe trên nền 4543 xe của Max-Pressure, tức dưới 1%. Vì vậy, chênh lệch ở chỉ số này cần được diễn giải ở mức vừa phải. Mức thay đổi tương đối lớn hơn

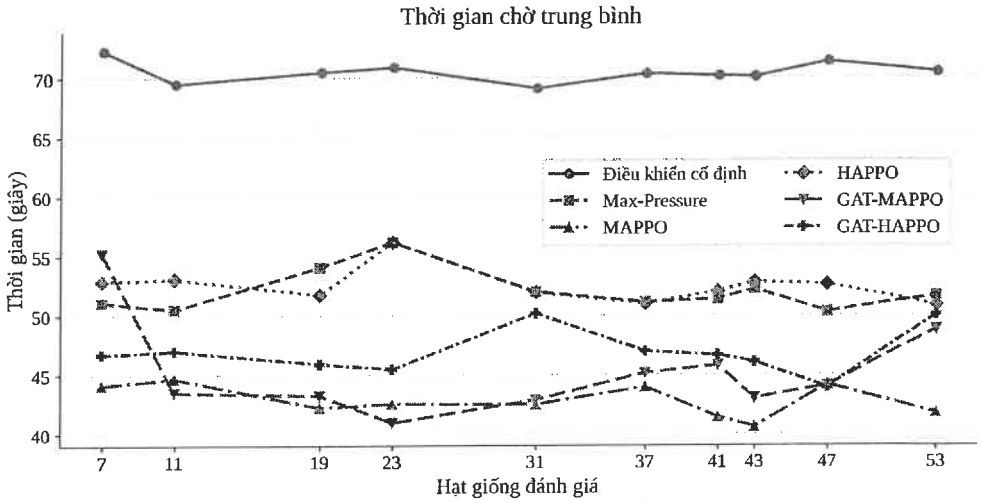
nằm ở thời gian chờ và tốc độ trung bình, đặc biệt đối với MAPPO.

HAPPO cho kết quả gần như ngang với Max-Pressure: số phương tiện hoàn thành chỉ cao hơn 1.1 xe, thời gian chờ cao hơn 0.46 giây và tốc độ thấp hơn 0.03 m/s. Kết quả này cho thấy việc sử dụng actor riêng không tự động tạo ra cải thiện trong kịch bản này. Một giả thuyết hợp lý là HAPPO có nhiều mạng chính sách riêng hơn, do đó có thể cần nhiều dữ liệu huấn luyện hoặc tinh chỉnh siêu tham số kỹ hơn để đạt hiệu quả ổn định. Ngược lại, MAPPO có thể tận dụng chia sẻ tham số giữa 41 tác tử, giúp kinh nghiệm từ nhiều nút giao cùng đóng góp vào một actor chung.

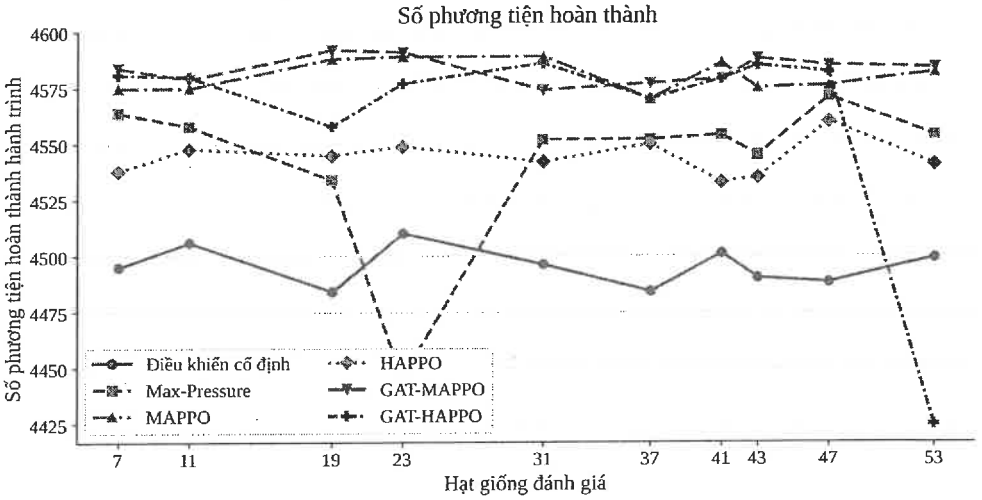
3.3.3 Vai trò của GAT

Kết quả cho thấy việc bổ sung GAT không tạo ra cải thiện đồng nhất trên toàn bộ chỉ số. Trong cặp MAPPO, GAT-MAPPO có số phương tiện hoàn thành hành trình cao nhất, nhưng mức chênh lệch so với MAPPO chỉ là 2.8 xe. Đồng thời, GAT-MAPPO có thời gian chờ cao hơn MAPPO 2.47 giây và tốc độ trung bình thấp hơn 0.09 m/s.

Trong cặp HAPPO, GAT-HAPPO giảm thời gian chờ trung bình so với HAPPO, từ 52.49 giây xuống 46.88 giây. Tuy nhiên, GAT-HAPPO vẫn không vượt MAPPO hoặc GAT-MAPPO ở các chỉ số chính, đồng thời có số sự kiện dịch chuyển bắt buộc trung bình cao nhất. Do đó, kết quả chỉ hỗ trợ kết luận thận trọng rằng GAT có thể giúp mô hình khai thác thêm quan hệ không gian, nhưng lợi ích phụ thuộc vào thuật toán nền và chỉ số đánh giá. Việc sử dụng GAT cũng cần được cân nhắc cùng với chi phí tính toán bổ sung, thay vì được xem như một cải tiến mặc định cho mọi biến thể thuật toán.



Hình 3.3: Thời gian chờ trung bình theo từng hạt giống đánh giá



Hình 3.4: Số phương tiện hoàn thành hành trình theo từng hạt giống đánh giá trên kịch bản Thanh Xuân UTM

Hình 3.4 cho thấy MAPPO và GAT-MAPPO duy trì số phương tiện hoàn thành hành trình ở mức cao và tương đối ổn định trên hầu hết các hạt giống đánh giá. Điều khiển cố định có độ dao động thấp nhưng số phương tiện hoàn thành hành trình thấp hơn rõ rệt. Một số phương pháp, đặc biệt là Max-Pressure và GAT-HAPPO, xuất hiện dao động lớn hơn ở một vài hạt giống, cho thấy cần xem xét kết quả trung bình cùng với độ lệch chuẩn thay vì chỉ dựa trên giá trị trung bình.

3.4 Diễn giải theo từng chỉ số

3.4.1 Thời gian chờ và tốc độ trung bình

Thời gian chờ trung bình là chỉ số thể hiện rõ nhất lợi ích của MAPPO. MAPPO đạt 42.79 giây, thấp nhất trong toàn bộ phương pháp. GAT-MAPPO đứng thứ hai với 45.26 giây, tiếp theo là GAT-HAPPO với 46.88 giây. Max-Pressure và HAPPO có thời gian chờ gần tương đương, lần lượt là 52.03 giây và 52.49 giây, trong khi điều khiển cố định kém nhất với 70.47 giây.

Tốc độ trung bình có xu hướng phù hợp với thời gian chờ. MAPPO đạt tốc độ cao nhất, 13.19 m/s, tiếp theo là GAT-MAPPO với 13.10 m/s. Điều này củng cố nhận định rằng MAPPO là phương pháp cân bằng nhất về chất lượng di chuyển trong kịch bản Thanh Xuân UTM.

3.4.2 Số phương tiện hoàn thành hành trình

GAT-MAPPO có số phương tiện hoàn thành hành trình cao nhất, 4583.3 xe, nhỉnh hơn MAPPO 2.8 xe và cao hơn Max-Pressure 40.3 xe. Tuy nhiên, do mức chênh lệch tương đối nhỏ, kết quả này không nên được diễn giải như ưu thế áp đảo. Có thể xem GAT-MAPPO là phương pháp có kết quả cao nhất ở chỉ số này, trong khi MAPPO tốt hơn nếu ưu tiên giảm thời gian chờ và tăng tốc độ trung bình.

3.4.3 Dịch chuyển bắt buộc

Số sự kiện dịch chuyển bắt buộc không cho thấy xu hướng cải thiện rõ ràng dưới các chính sách học tăng cường. Điều khiển cố định có giá trị trung bình thấp nhất, 8.1 sự kiện, trong khi MAPPO là 8.4, GAT-MAPPO là 9.5, HAPPO là 9.6, Max-Pressure là 9.9 và GAT-HAPPO là 10.0. Tuy nhiên, điều khiển cố định lại có thời gian chờ và tốc độ kém nhất, cho thấy chỉ số này không phản ánh trực tiếp chất lượng điều khiển trong kịch bản hiện tại.

Vì vậy, số sự kiện dịch chuyển bắt buộc chỉ được sử dụng như một chỉ số phụ để theo dõi mức độ bế tắc trong mô phỏng, không phải bằng chứng chính để kết luận phương pháp nào tốt hơn.

3.5 Nhận xét kết quả

Kết quả thực nghiệm cho thấy các phương pháp học tăng cường dựa trên PPO cải thiện rõ rệt so với điều khiển cố định ở các chỉ số vận hành chính trong kịch bản Thanh Xuân UTM. So với Max-Pressure, MAPPO và GAT-MAPPO đạt số phương tiện hoàn thành hành trình và tốc độ trung bình cao hơn, đồng thời giảm thời gian chờ trung bình. Tuy nhiên, số sự kiện dịch chuyển bắt buộc không cho thấy xu hướng

cải thiện nhất quán, nên chỉ số này không được sử dụng làm căn cứ chính để kết luận phương pháp nào tốt hơn.

Nếu xét chất lượng di chuyển tổng thể, MAPPO là phương pháp có kết quả cân bằng nhất trong kịch bản đánh giá. Phương pháp này đạt thời gian chờ trung bình thấp nhất, tốc độ trung bình cao nhất và số phương tiện hoàn thành hành trình gần tương đương GAT-MAPPO. Nếu chỉ xét riêng số phương tiện hoàn thành hành trình, GAT-MAPPO là phương pháp đạt giá trị cao nhất. Tuy nhiên, mức chênh lệch giữa GAT-MAPPO và MAPPO ở chỉ số này tương đối nhỏ, nên không nên diễn giải đây là một ưu thế áp đảo.

Kết quả cũng cho thấy việc bổ sung GAT không bảo đảm cải thiện đồng nhất trên mọi chỉ số. GAT-MAPPO nhỉnh hơn MAPPO về số phương tiện hoàn thành hành trình, nhưng MAPPO lại tốt hơn về thời gian chờ trung bình và tốc độ trung bình. Trong khi đó, HAPPO không thể hiện ưu thế rõ ràng so với Max-Pressure. Kết quả hiện tại vì vậy chưa đủ để kết luận rằng chính sách riêng hoặc cơ chế truyền tin đồ thị luôn mang lại lợi ích trong kịch bản Thanh Xuân UTM.

Các nhận xét trên cho thấy tầm quan trọng của việc sử dụng Max-Pressure làm phương pháp đối chứng thích ứng. Nếu chỉ so sánh với điều khiển cố định, các thuật toán học tăng cường có thể tạo ra mức cải thiện lớn ở thời gian chờ và tốc độ trung bình. Tuy nhiên, khi so sánh với Max-Pressure, hiệu quả của từng thuật toán cần được phân tích cẩn thận theo từng chỉ số thay vì kết luận một phương pháp vượt trội trên toàn bộ tiêu chí.

3.6 Hạn chế

Nghiên cứu vẫn tồn tại một số hạn chế cần được xem xét khi diễn giải kết quả. Trước hết, toàn bộ kết quả hiện tại chỉ được kiểm chứng trong môi trường mô phỏng SUMO và chưa được đánh giá ngoài thực địa. Vì vậy, các kết luận của đề án phản ánh hiệu quả trong phạm vi kịch bản mô phỏng Thanh Xuân UTM, chưa thể được xem là bằng chứng trực tiếp cho khả năng triển khai thực tế.

Bên cạnh đó, mạng lưới giao thông được xây dựng từ dữ liệu OSM, nên có thể tồn tại sai lệch so với hình học đường, số làn hoặc quy tắc rẽ thực tế tại một số vị trí. Nhu cầu giao thông cũng được sinh từ dữ liệu khảo sát và điều chỉnh về mức mục tiêu, chưa phải là ma trận nhu cầu đầy đủ giữa các điểm đi và điểm đến trong toàn bộ khu vực. Những yếu tố này có thể ảnh hưởng đến độ chính xác của mô phỏng và kết quả đánh giá.

Một hạn chế khác là chương này mới tập trung vào một mức nhu cầu chính. Việc đánh giá thêm các mức tải thấp hơn và cao hơn sẽ giúp kiểm tra tốt hơn khả năng tổng quát của chính sách. Ngoài ra, các biến thể HAPPO và GAT có thể nhạy với số lần mô phỏng huấn luyện, thứ tự cập nhật, cấu trúc đồ thị và siêu tham số huấn luyện. Do đó, kết quả hiện tại chưa chứng minh rằng các cấu hình này đã đạt tối ưu.

Cuối cùng, số sự kiện dịch chuyển bắt buộc có dao động giữa các hạt giống và chịu ảnh hưởng từ cấu hình mạng đường. Vì vậy, đề án chỉ sử dụng chỉ số này như một thông tin hỗ trợ để theo dõi mức độ bế tắc trong mô phỏng, thay vì xem đây là bằng chứng chính về hiệu quả điều khiển.

3.7 Kết luận chương

Chương này đã trình bày kịch bản đánh giá Thanh Xuân UTM, các phương pháp so sánh, hệ thống chỉ số và kết quả định lượng trên 10 hạt giống. Xét theo các chỉ số vận hành chính gồm số phương tiện hoàn thành hành trình, thời gian chờ trung bình và tốc độ trung bình, MAPPO và GAT-MAPPO là hai phương pháp có kết quả tốt nhất trong nhóm thuật toán được đánh giá. MAPPO đạt thời gian chờ trung bình thấp nhất và tốc độ trung bình cao nhất, trong khi GAT-MAPPO có số phương tiện hoàn thành hành trình cao nhất. Tuy nhiên, không có phương pháp nào vượt trội trên toàn bộ chỉ số; các biến thể GAT không luôn cải thiện so với biến thể không dùng GAT, và số sự kiện dịch chuyển bắt buộc không cho thấy xu hướng cải thiện rõ ràng. Do đó, kết quả cần được diễn giải theo từng chỉ số, với Max-Pressure đóng vai trò là phương pháp đối chứng thích ứng quan trọng.

KẾT LUẬN

Các kết quả đạt được của đề án

Trải qua quá trình nghiên cứu và thực hiện, đề án đã hoàn thành các mục tiêu đặt ra trong phạm vi mô phỏng. Kết quả đầu tiên của đề án là xây dựng được một quy trình chuyển đổi từ dữ liệu khảo sát UTM-Hanoi sang kịch bản mô phỏng vi mô SUMO cho khu vực Thanh Xuân, Hà Nội. Quy trình này bao gồm việc xây dựng mạng đường từ dữ liệu bản đồ, sinh nhu cầu giao thông hỗn hợp gồm xe máy, ô tô và xe buýt, đồng thời sử dụng hệ số PCU để quy đổi các loại phương tiện về cùng một đơn vị đo trong quá trình trích xuất trạng thái và tính toán chỉ số.

Bên cạnh quy trình mô phỏng, đề án đã đặc tả một môi trường điều khiển đèn tín hiệu đa nút giao phục vụ huấn luyện và đánh giá các tác tử học tăng cường. Môi trường này xác định rõ tập tác tử điều khiển, vector quan sát cục bộ dựa trên PCU, không gian hành động có cơ chế lọc hành động hợp lệ và các ràng buộc chuyển pha an toàn. Ngoài ra, đề án xây dựng giao thức lưu mô hình kèm bản ghi cấu hình, qua đó hạn chế việc tiếp tục huấn luyện hoặc đánh giá một bản lưu mô hình trên một cấu hình mô phỏng không tương thích.

Trên môi trường đã xây dựng, đề án triển khai và so sánh bốn thuật toán học tăng cường đa tác tử gồm MAPPO, HAPPO, GAT-MAPPO và GAT-HAPPO trong cùng một mức nhu cầu và cùng hệ thống chỉ số đánh giá. Các biến thể GAT khai thác quan hệ lân cận giữa các nút giao thông qua đồ thị chính sách được xây dựng từ mạng SUMO, trong khi các biến thể không dùng GAT xử lý quan sát cục bộ bằng MLP. Cách tổ chức này cho phép đề án đánh giá ảnh hưởng của hai yếu tố chính: chính sách dùng chung hoặc chính sách riêng, và có hoặc không có cơ chế truyền tin đồ thị.

Kết quả đánh giá định lượng được thực hiện trên nhiều hạt giống ngẫu nhiên với bốn chỉ số gồm số phương tiện hoàn thành hành trình, thời gian chờ trung bình, tốc độ trung bình và số sự kiện dịch chuyển bắt buộc. Xét theo các chỉ số vận hành chính, MAPPO và GAT-MAPPO đạt kết quả tốt trong nhóm phương pháp được đánh giá. Tuy nhiên, kết quả cũng cho thấy không có phương pháp nào vượt trội trên toàn bộ tiêu chí. Vì vậy, các kết quả của đề án cần được diễn giải theo từng chỉ số cụ thể và trong phạm vi kịch bản Thanh Xuân UTM đã xác định.

Nhận xét, đề xuất, khuyến nghị

Quá trình nghiên cứu cho thấy việc áp dụng MARL vào điều khiển giao thông đô thị Việt Nam là khả thi về mặt kỹ thuật trong môi trường mô phỏng. Tuy nhiên, kết quả thực nghiệm cũng cho thấy cần thận trọng khi đánh giá hiệu quả của các thuật toán học tăng cường. Max-Pressure là một phương pháp điều khiển thích ứng mạnh và không nên được xem là một đối chứng yếu. Vì vậy, các mô hình học tăng cường cần được so sánh không chỉ với điều khiển cố định mà còn với Max-Pressure để làm rõ mức cải thiện thực sự.

Chất lượng dữ liệu đầu vào cũng có ảnh hưởng đáng kể đến độ tin cậy của kết quả. Mạng lưới được xây dựng từ OSM và nhu cầu giao thông được sinh từ dữ liệu khảo sát giúp kịch bản mô phỏng gần với bối cảnh nghiên cứu hơn so với mạng tổng hợp. Tuy nhiên, các yếu tố như chiều rộng làn đường, cấu trúc hình học chi tiết của nút giao và ma trận nhu cầu giữa các điểm đi và điểm đến vẫn cần được cải thiện nếu muốn tăng độ chính xác của mô phỏng.

Về mặt triển khai, các thuật toán MARL trong đề án vẫn đòi hỏi chi phí huấn luyện đáng kể. Khi mở rộng sang mạng lưới lớn hơn hoặc nhiều mức nhu cầu hơn, hệ thống cần được hỗ trợ bởi hạ tầng tính toán phù hợp, chẳng hạn GPU hoặc cơ chế xử lý song song. Do đó, việc ứng dụng các mô hình này ngoài thực tế cần được xem xét như một hướng nghiên cứu tiếp theo, thay vì là kết luận trực tiếp từ phạm vi mô phỏng hiện tại.

Hướng nghiên cứu tiếp theo

Trong tương lai, nghiên cứu có thể được mở rộng bằng cách đánh giá các thuật toán trên nhiều mức nhu cầu giao thông khác nhau. Việc bổ sung các kịch bản nhu cầu thấp hơn và cao hơn sẽ giúp kiểm tra độ ổn định của chính sách, đồng thời làm rõ hơn quan hệ so sánh giữa các thuật toán học tăng cường và Max-Pressure trong các điều kiện vận hành khác nhau. Bên cạnh đó, việc tăng số hạt giống ngẫu nhiên và báo cáo khoảng tin cậy sẽ giúp kết quả thống kê có độ tin cậy cao hơn.

Một hướng phát triển khác là cải thiện thiết kế hàm phần thưởng và cách xử lý các hiện tượng bế tắc trong mô phỏng. Thành phần phạt liên quan đến dịch chuyển bắt buộc cần được nghiên cứu thêm để giảm các trường hợp ùn tắc nghiêm trọng mà không làm suy giảm các chỉ số vận hành khác như thời gian chờ hoặc tốc độ trung bình. Hướng này đặc biệt quan trọng trong các mạng lưới dày đặc, nơi ùn tắc tại một nút có thể lan truyền nhanh sang các nút lân cận.

Ngoài ra, nghiên cứu trong tương lai cần tiếp tục cải thiện chất lượng dữ liệu giao thông đầu vào. Việc bổ sung dữ liệu chiều rộng làn đường, cấu trúc hình học chi tiết của nút giao và ma trận nhu cầu giữa các điểm đi và điểm đến từ dữ liệu thực địa sẽ giúp kịch bản mô phỏng đáng tin cậy hơn. Nếu có dữ liệu lịch trình tuyến buýt thực tế, mô hình cũng có thể mô phỏng xe buýt chính xác hơn và xem xét thêm yếu tố ưu tiên xe buýt trong hàm phần thưởng.

TÀI LIỆU THAM KHẢO

[1] Pablo Alvarez Lopez et al. (2018) "Microscopic Traffic Simulation using SUMO", *The 21st IEEE International Conference on Intelligent Transportation Systems*.

[2] F.V. Webster (1958). *Traffic Signal Settings*, H.M. Stationery Office.

[3] Pravin Varaiya (2013) "Max pressure control of a network of signalized intersections", *Transportation Research Part C: Emerging Technologies*, 36, 177–195.

[4] Hua Wei et al. (2019) "PressLight: Learning Max Pressure Control to Coordinate Traffic Signals in Arterial Network", *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1290–1298.

[5] Volodymyr Mnih et al. (2015) "Human-level control through deep reinforcement learning", *Nature*, 518 (7540), 529–533.

[6] Hua Wei et al. (2018) "IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control", *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2496–2505.

[7] Guanjie Zheng et al. (2019) "Learning Phase Competition for Traffic Signal Control", *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 196–205.

[8] Hua Wei et al. (2019) "CoLight: Learning Network-level Cooperation for Traffic Signal Control", *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1913–1922.

[9] Chacha Chen et al. (2020) "Toward a Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control", *Proceedings of the 2020 Conference on Innovative Applications of Artificial Intelligence*.

[10] John Schulman et al. (2017) "Proximal Policy Optimization Algorithms", *CoRR*, abs/1707.06347.

[11] Ronald J. Williams (1992) "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning", *Machine Learning*, 8, 229–256.

[12] Aicha Saadi et al. (2025) "A survey of reinforcement and deep reinforcement learning for coordination in intelligent traffic light control", *Journal of Big Data*, 12 (1).

[13] Michael L. Littman (1994) "Markov games as a framework for multi-agent reinforcement learning", *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, 157–163.

[14] Ming Tan (1993) "Multi-agent reinforcement learning: independent versus cooperative agents", *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, 330–337.

[15] Ryan Lowe et al. (2017) "Multi-agent actor-critic for mixed cooperative-competitive environments", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6382–6393.

[16] Tabish Rashid et al. (2018) "QMIX: Monotonic Value Factorisation for Deep Multi-Agent Reinforcement Learning", *arXiv preprint arXiv:1803.11485.*

[17] Chao Yu et al. (2021) "The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games", *arXiv preprint arXiv:2103.01961.*

[18] Yifan Zhong et al. (2024) "Heterogeneous-Agent Reinforcement Learning", *Journal of Machine Learning Research*, 25 (32), 1–67.

[19] Jing Shang et al. (2025) "Graph-Based Cooperation Multiagent Reinforcement Learning for Intelligent Traffic Signal Control", *IEEE Internet of Things Journal*, 12 (10), 14362–14374.

[20] Thomas N Kipf and Max Welling (2017) "Semi-Supervised Classification with Graph Convolutional Networks", *arXiv preprint arXiv:1609.02907.*

[21] Petar Veličković et al. (2017) "Graph Attention Networks".

[22] Bing Yu, Haoteng Yin, and Zhanxing Zhu (2018) "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting", *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3634–3640.

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 1949/QĐ-HV ngày 07 tháng 07 năm 2026 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi...8....giờ..00..phút, ngày..22.. tháng 07 năm 2026 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Nguyễn Thanh Hùng**

Tên đề án tốt nghiệp: **Nghiên cứu mô hình học tăng cường đa tác tử trong bài toán điều khiển tín hiệu giao thông tại Việt Nam**

Chuyên ngành: **Hệ thống thông tin** Mã số: **8480104**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ..05../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**PGS.TS. Nguyễn Trọng Khánh**	Chủ tịch	
2	**TS. Đặng Ngọc Hùng**	Thư ký	
3	**TS. Trịnh Bảo Ngọc**	Phản biện 1	
4	**TS. Nguyễn Văn Vinh**	Phản biện 2	
5	**TS. Vũ Văn Thỏa**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên. (có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

1. Cơ chế cuả GAT như thế nào trong bài toán đa tác tử
2. Tại sao lại dùng mô hình học tăng cường đa tác tử trong bài toán điều khiển tín hiệu giao thông
3. Tại sao GAT phù hợp với thuật toán đề xuất mà bài chỉ cuả NO ?
4. Có những loại tác tử nào trong áp dụng

7. Trả lời của học viên:

1, Học viên có đã lời và giải thích cô thể hoạt động nhỏ GAT trong bài toán

2, Học viên có giải thích lý do lựa chọn đa tác thì đa thì liên kết, mối liên quan các mút / đỉnh giao thông

3, Học viên chưa làm rõ và giải thích được lý do GAT phù hợp thuật toán nên và lựa chọn thuật toán PPO.

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: PGS TS Nguyễn Trang Khánh
  2. Uỷ viên Ban kiểm phiếu: TS Vũ Văn Thoa
  3. Uỷ viên Ban kiểm phiếu: TS Đào Ngọc Hùng
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 8.0

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

Rà soát chỉnh sửa nội dung về trình bày và văn phong

Nhờ góp ý của thầy cô

Chuẩn hoá tài liệu tham khảo, danh mục, hàng, hình

Thuật ngữ

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: Đề nghị công nhận

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS: Không

Buổi làm việc kết thúc vào..... cùng ngày.

Chủ tịch	Thư ký
Nguyễn Trang Khánh	Đào Ngọc Hùng

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: NGHIÊN CỨU MÔ HÌNH HỌC TĂNG CƯỜNG ĐA TÁC TỬ TRONG BÀI TOÁN ĐIỀU KHIỂN TÍN HIỆU GIAO THÔNG TẠI VIỆT NAM

Chuyên ngành: Hệ thống thông tin

Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: Nguyễn Thanh Hùng .....

Họ và tên người nhận xét: Nguyễn Văn Vinh

Học hàm, học vị: Tiến sỹ .....

Chuyên ngành: Khoa học máy tính .....

Cơ quan công tác: Trường Đại học Công Nghệ, ĐHQG Hà Nội .....

Số điện thoại:.....E-mail:.....

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:
Việt Nam là nước có tốc độ đô thị hóa nhanh và phương tiện giao thông cá nhân tăng là thách thức cho hệ thống giao thông. Gần đây việc sử dụng học tăng cường đa tác tử có nhiều ứng dụng trong thực tế đặc biệt đối với bài toán toán đa tác tử. Vì vậy đề tài nghiên cứu và ứng dụng học tăng cường đa tác tử cho bài toán điều khiển tín hiệu giao thông là có ý nghĩa về mặt khoa học cũng như thực tế.
II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

- Trình bày tổng quan về học tăng cường trong điều khiển giao thông
- Nghiên cứu và xây dựng mô hình học tăng cường (triên kha và thử nghiệm với 4 thuật toán) trong điều khiển giao thông tại Việt Nam
- Thử nghiệm và đánh giá với kịch bản Thanh Xuân UTM cho thấy thuật toán MAPPO và GAT-MAPPO đạt kết quả tốt hơn trong các thuật toán được đánh giá.

III/ Những vấn đề cần giải thích thêm:

- Vì đây là 1 bài toán phức tạp, nên tập trung vào 1 thuật toán và ta thử nghiệm và đánh giá kỹ thì tốt hơn là 4 thuật toán nhưng còn khá đơn giản?

- Vấn đề các siêu tham số. Bảng 2.2 cho thấy một bộ siêu tham số duy nhất được áp dụng cho cả 4 thuật toán. Đây là vấn đề lớn vì HAPPO có kiến trúc phức tạp hơn hẳn (41 actor riêng thay vì 1 actor chung) → cần learning rate, batch size, số epoch khác biệt để hội tụ ổn định. Kết quả HAPPO gần như ngang Max-Pressure hoàn toàn có thể là do thiếu tuning chứ không phải do bản chất thuật toán kém? So sánh 4 thuật toán mà không kiểm soát được liệu mỗi thuật toán đã đạt hiệu năng tối đa của nó hay chưa, thì việc kết luận "MAPPO tốt hơn HAPPO" cần được xem xét kỹ so với vẻ ngoài của bảng số liệu?
- Cỡ mẫu huấn luyện rất nhỏ so với độ phức tạp bài toán. 360,000 quyết định bước (500 episode × 720 bước) cho một hệ 41 tác tử với không gian hành động biến thiên là con số khá nhỏ.
- Toàn bộ kết luận dựa trên một topology mạng lưới duy nhất (Thanh Xuân, 41 tác tử). Đây là hạn chế phổ biến trong MARL traffic nhưng cần nêu rõ: không có bằng chứng nào cho thấy policy học được có transfer sang mạng lưới khác, hoặc thậm chí cùng mạng lưới nhưng cấu trúc demand thay đổi?
- Vai trò của GAT chưa được phân tích đủ sâu để rút ra kết luận. Đề án kết luận "lợi ích của GAT phụ thuộc vào thuật toán nền", tại sao?
- Một số lỗi chính tả và thuật ngữ cần được chỉnh sửa. Ví dụ như "cưỡng bức" (tr. 53).

IV/ Kết luận:

Đề án đáp ứng cơ bản đề án bậc thạc sỹ ngành hệ thống thông tin sau khi chỉnh sửa theo các góp ý của các thành viên hội đồng.

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày...20..tháng...7...năm ..2026

NGƯỜI NHẬN XÉT

Nguyễn Văn Vinh

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: Nghiên cứu mô hình học tăng cường đa tác tử trong bài toán điều khiển tín hiệu giao thông tại Việt Nam

Chuyên ngành: Hệ thống thông tin
Mã chuyên ngành: 8.48.01.04
Họ và tên học viên: Nguyễn Thanh Hùng
Họ và tên người nhận xét: Trịnh Bảo Ngọc
Học hàm, học vị: Tiến sĩ
Chuyên ngành: Kỹ thuật phần mềm
Cơ quan công tác: Trường đại học Hà Nội
Số điện thoại: 0987 355 799 E-mail: ngoctb@hanu.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Cơ sở khoa học: Đề tài được xây dựng trên nền tảng khoa học của học tăng cường (Reinforcement Learning - RL), học tăng cường đa tác tử (Multi-Agent Reinforcement Learning - MARL) và mạng nơ-ron đồ thị (Graph Neural Network - GNN). Tác giả đã phân tích các phương pháp điều khiển tín hiệu truyền thống như chu kỳ cố định, điều khiển theo nhu cầu, áp lực cực đại, đồng thời so sánh với các tiếp cận hiện đại dựa trên học máy, học sâu và học tăng cường. Việc kết hợp MARL với cơ chế chú ý đồ thị (Graph Attention Network - GAT) cho thấy cơ sở khoa học rõ ràng, phù hợp với xu hướng nghiên cứu tiên tiến trong điều khiển giao thông đô thị.

Cơ sở thực tiễn: Đề tài xuất phát từ thực trạng giao thông đô thị tại Việt Nam, đặc biệt là Hà Nội, nơi tỷ lệ xe máy cao và hành vi di chuyển không tuân thủ làn đường nghiêm ngặt. Các hệ thống điều khiển tín hiệu hiện nay chủ yếu dựa trên chu kỳ cố định, thiếu khả năng thích ứng với biến động thực tế. Tác giả đã xây dựng môi trường mô phỏng vi mô trên nền tảng SUMO, sử dụng dữ liệu khảo sát UTM và bản đồ OSM để phản ánh đặc thù giao thông hỗn hợp. Việc thử nghiệm các thuật toán MARL trong kịch bản Thanh Xuân cho thấy tính thực tiễn cao, có thể hỗ trợ cải thiện hiệu quả điều khiển tín hiệu giao thông.

Tính cấp thiết: Trong bối cảnh đô thị hóa nhanh và lượng phương tiện cá nhân tăng mạnh, ùn tắc giao thông trở thành vấn đề nghiêm trọng tại các thành phố lớn ở Việt Nam. Các phương pháp điều khiển tín hiệu truyền thống không còn đáp ứng được nhu cầu thực tế. Do đó, việc nghiên cứu mô hình học tăng cường đa tác tử cho điều khiển tín hiệu giao thông là cấp thiết, nhằm nâng cao khả năng thích ứng, giảm thời gian

chờ, tăng tốc độ trung bình và cải thiện chất lượng sống đô thị. Đây là nhu cầu thực tế và cấp bách đối với hệ thống giao thông Việt Nam hiện nay.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Kết cấu đề án:

Đề án gồm ba chương trong 71 trang bao gồm: tổng quan về học tăng cường trong điều khiển tín hiệu giao thông; mô hình học tăng cường đa tác tử và kiến trúc thuật toán; thực nghiệm và đánh giá kết quả trên kịch bản Thanh Xuân UTM. Cuối cùng là phần kết luận và tài liệu tham khảo.

Tóm tắt các kết quả đạt được:

Đề tài đã đạt được nhiều kết quả quan trọng. Về lý luận, tác giả đã hệ thống hóa kiến thức nền tảng về học tăng cường, MARL và GNN, đồng thời phân tích khoảng trống nghiên cứu trong bối cảnh giao thông hỗn hợp tại Việt Nam. Về thực tiễn, tác giả đã xây dựng môi trường mô phỏng vi mô trên SUMO cho khu vực Thanh Xuân, sử dụng dữ liệu UTM và OSM để phản ánh đặc thù giao thông hỗn hợp.

Trên môi trường này, đề tài đã đặc tả bài toán điều khiển tín hiệu dưới dạng hệ đa tác tử, thiết kế không gian quan sát, hành động, hàm phần thưởng và cơ chế phối hợp. Bốn thuật toán MARL gồm MAPPO, HAPPO, GAT-MAPPO và GAT-HAPPO đã được triển khai và so sánh với các phương pháp đối chứng như điều khiển cố định và Max-Pressure.

Kết quả thực nghiệm cho thấy các thuật toán MARL, đặc biệt là GAT-MAPPO và GAT-HAPPO, có khả năng giảm thời gian chờ, tăng số phương tiện hoàn thành hành trình và cải thiện tốc độ trung bình. Đây là bằng chứng thực nghiệm quan trọng về hiệu quả của học tăng cường đa tác tử trong điều khiển tín hiệu giao thông tại Việt Nam.

III/ Những vấn đề cần giải thích thêm:

Về hình thức trình bày:

Đề án trình bày rõ ràng, có mục lục chi tiết, danh mục thuật ngữ và hình vẽ đầy đủ, giúp người đọc dễ theo dõi và nắm bắt nội dung.

Một số lỗi định dạng hoặc chính tả xuất hiện nhiều trong tài liệu cần được chỉnh sửa. Trang ii, mục "2.3.3 Không gian hành động và cơ chế lọc hành động" bị ghi nhầm thành "3.3", gây nhầm lẫn. Trang iv, bảng từ viết tắt có lỗi căn chỉnh khi một số thuật ngữ bị ngắt dòng không hợp lý, ví dụ "Application Programming Interface" bị tách thành "In-terface".

Trang v, bảng từ viết tắt có lỗi trình bày khi "SUMO" được ghi kèm "[1]" không nhất quán với các thuật ngữ khác. Trang vii, danh sách bảng có lỗi căn chỉnh khi số trang

không thẳng hàng. Trang viii, danh sách hình vẽ có lỗi xuống dòng không hợp lý, ví dụ "Hình 3.1 Mạng lưới giao thông khu vực Thanh Xuân" bị tách dòng gây khó đọc.

Về nội dung:

Một số vấn đề về kiến thức và logic cần lưu ý. Trang 5, phần điều khiển chu kỳ cố định trình bày công thức Webster nhưng chưa giải thích rõ ý nghĩa các tham số trong bối cảnh giao thông hỗn hợp tại Việt Nam, khiến người đọc khó hình dung tính ứng dụng.
Trang 12, phần khung lý thuyết Markov đa tác tử chưa phân tích sâu sự khác biệt giữa MDP đơn tác tử và trò chơi Markov đa tác tử, chỉ dừng ở mức khái quát.
Trang 22, phần mô hình hóa bài toán chưa giải thích rõ cách xây dựng không gian quan sát từ dữ liệu thực tế, chỉ mô tả tổng quát.
Trang 26, phần hàm phần thưởng chưa phân tích chi tiết tác động của từng thành phần lên kết quả huấn luyện, khiến người đọc khó đánh giá tính hợp lý. Trang 39, bảng thông số huấn luyện chưa giải thích rõ lý do chọn tham số, chỉ liệt kê.
Trang 46, bảng kết quả trung bình ± độ lệch chuẩn chưa có so sánh định lượng cụ thể với các phương pháp đối chứng, chỉ dừng ở mức mô tả. Trang 50, phần diễn giải theo từng chỉ số chưa phân tích sâu mối quan hệ giữa thời gian chờ, tốc độ trung bình và số phương tiện hoàn thành hành trình. Trang 51, phần hạn chế chưa đề cập đến chi phí tính toán và khả năng triển khai thực tế ngoài mô phỏng, yếu tố quan trọng trong ứng dụng. Những khuyết điểm này cho thấy cần bổ sung phân tích sâu hơn về tham số kỹ thuật, dữ liệu thực nghiệm và khả năng triển khai thực tế để tăng tính toàn diện và giá trị ứng dụng..

Câu hỏi:

1. Cơ chế GAT cải thiện phối hợp giữa các nút giao như thế nào?
2. Vì sao MARL phù hợp hơn RL đơn tác tử trong điều khiển tín hiệu giao thông?
3. Vai trò của hàm phần thưởng trong huấn luyện MARL là gì?

IV/ Kết luận:
Đồng ý.

Ngày 21 tháng 07 năm 2026
NGƯỜI NHẬN XÉT

TS Trịnh Bảo Ngọc

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP

Họ và tên học viên: Nguyễn Thanh Hùng
Chuyên ngành: Hệ thống thông tin
Khóa: 2024 đợt 2
Tên đề tài: Nghiên cứu mô hình học tăng cường đa tác tử trong bài toán điều khiển tín hiệu giao thông tại Việt Nam
Người hướng dẫn khoa học: TS. Nguyễn Đình Hoá
Ngày bảo vệ: 22/07/2026

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Rà soát, chỉnh sửa nội dung về trình bày văn phong như góp ý của Hội đồng.	Học viên đã thay thuật ngữ "dịch chuyển cưỡng bức" bằng "dịch chuyển bắt buộc" trong toàn bộ đề án. Đồng thời, học viên đã rà soát, chỉnh sửa các lỗi soạn thảo trong toàn bộ đề án.
2	Chuẩn hóa tài liệu tham khảo, danh mục, bảng biểu, thuật ngữ.	Học viên đã chuyển trích dẫn tài liệu tham khảo của SUMO từ danh mục chữ viết tắt vào lần đầu giới thiệu SUMO trong nội dung đề án. Đồng thời, học viên đã rà soát, chuẩn hóa tài liệu tham khảo, danh mục, bảng biểu và thuật ngữ theo quy định trình bày đề án.

Hà Nội, ngày 28 tháng 7 năm 2026

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
PGS.TS. Nguyễn Trọng Khánh	TS. Đặng Ngọc Hùng	TS. Nguyễn Đình Hoá	Nguyễn Thanh Hùng

13% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm cả các nguồn trùng lặp, cho mỗi c...

Đã lọc khỏi Báo cáo

- Mục lục tham khảo
- Văn bản được trích dẫn
- Văn bản được trích dẫn
- Kết quả trùng khớp nhỏ (ít hơn 8 từ)

Nguồn hàng đầu

10% Nguồn Internet
4% Ấn bản
4% Bài tập được nộp (bài của học sinh)

Học viên

Hùng

Nguyễn Thanh Hùng

Người hướng dẫn khoa học

Nguyễn Đình Hoá