BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH

Nguyễn Nhựt Đăng
Ngô Trọng Nghĩa

Xây dựng hệ thống mô phỏng dự báo và điều tiết giao thông dựa trên dữ liệu snapshot từ camera giao thông sử dụng Học sâu và mô phỏng SUMO

KHÓA LUẬN TỐT NGHIỆP

NGÀNH CÔNG NGHỆ KỸ THUẬT MÁY TÍNH

TP. Hồ Chí Minh, Tháng 01, 2026

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH

Nguyễn Nhựt Đăng
Ngô Trọng Nghĩa

Xây dựng hệ thống mô phỏng dự báo và điều tiết giao thông dựa trên dữ liệu snapshot từ camera giao thông sử dụng Học sâu và mô phỏng SUMO

NGÀNH CÔNG NGHỆ KỸ THUẬT MÁY TÍNH

Người hướng dẫn khoa học: TS. Huỳnh Thế Thiện

TP. Hồ Chí Minh, Tháng 01, 2026

width=!,height=!,pages=-, noautoscale=true, width=height=

Lời Cam Đoan

Với tư cách là người thực hiện khóa luận tốt nghiệp này, chúng tôi là Ngô Trọng Nghĩa, mã số sinh viên 21161155 và Nguyễn Nhựt Đăng, mã số sinh viên 21119062 cùng đang theo học ngành Công nghệ kỹ thuật máy tính tại Khoa Điện - Điện tử, Trường Đại học Sư phạm Kỹ thuật TP.HCM. Chúng tôi xin khẳng định đây hoàn toàn là công trình nghiên cứu do chúng tôi sáng tạo. Nội dung và kết quả của khóa luận phản ánh năng lực chuyên môn, kỹ năng nghiên cứu và sự nỗ lực tự thân của chúng tôi, không hề vay mượn hay sao chép từ bất kỳ đồ án, bài báo hay tài liệu nào đã được công bố trước đây mà không trích dẫn nguồn. Chúng tôi cam đoan mọi tài liệu tham khảo được sử dụng đều đã được ghi nhận đầy đủ và chính xác, tuân thủ nghiêm ngặt quy định về trích dẫn của Nhà trường và các chuẩn mực học thuật quốc tế. Chúng tôi đảm bảo tính xác thực, khách quan của thông tin trình bày và khẳng định không có hành vi học thuật không trung thực. Chúng tôi hoàn toàn chịu trách nhiệm về tính nguyên bản của công trình này và chấp nhận mọi hình thức xử lý kỷ luật nếu phát hiện bất kỳ vi phạm nào đối với bản cam kết này.

Người thực hiện tiểu luận
(Ký và ghi rõ họ tên)

Lê Trường Thịnh

Lời Cảm Tạ

Chúng tôi nhận thức sâu sắc rằng việc hoàn thành khóa luận tốt nghiệp này không thể thực hiện được nếu thiếu đi sự đồng hành, tư vấn và hỗ trợ quý báu từ quý Thầy Cô cùng các bạn bè trong ngành Hệ thống nhúng và IoT, Khoa Điện - Điện tử, Trường Đại Học Sư Phạm Kỹ Thuật Thành phố Hồ Chí Minh. Chúng tôi xin bày tỏ lòng biết ơn chân thành nhất đến tất cả những người đã dành thời gian, công sức góp ý và giúp đỡ chúng tôi trong suốt quá trình thực hiện công trình này. Đặc biệt, sự dẫn dắt của Thầy Huỳnh Thế Thiện đóng vai trò then chốt. Những định hướng chuyên môn, lời khuyên tận tình và lộ trình nghiên cứu Thầy vạch ra ngay từ những bước đi đầu tiên đã ảnh hưởng sâu sắc đến tư duy và cách tiếp cận đề tài của chúng tôi. Dù đã rất cố gắng, khóa luận chắc chắn vẫn còn tồn tại những điểm chưa hoàn thiện. Chúng tôi rất mong nhận được những góp ý thẳng thắn và đánh giá khách quan để có thể nâng cao kiến thức, khắc phục hạn chế và tạo ra những sản phẩm chất lượng hơn trong tương lai. Xin chân thành cảm ơn!

Tóm Tắt

Nghiên cứu này giới thiệu một phương pháp mới dựa trên mô hình học sâu để nhận diện tín hiệu 5G (fifth-generation), còn được gọi là NR (new radio), và LTE (long-term evolution), với trọng tâm là xác định các vùng phổ tần số của tín hiệu được điều chế trong mạng vô tuyến. Phương pháp này nhằm mục đích hỗ trợ việc xây dựng các mạng vô tuyến nhận thức thế hệ tiếp theo. Về mặt lý thuyết, trong quá trình truyền dẫn, các tín hiệu được điều chế thường trở nên khó nhận diện do có dạng sóng mang phức tạp. Để giải quyết vấn đề này, các tín hiệu thu được từ máy thu sẽ được chuyển đổi thành hình ảnh phổ, giúp hiển thị thông tin trực quan hơn bằng cách áp dụng phép biến đổi Fourier thời gian ngắn (short-time Fourier transform - STFT).

Để xác định vùng phổ của tín hiệu 5G và LTE cùng tồn tại trên một hình ảnh phổ, tác giả giới thiệu một phương pháp cảm biến phổ tiên tiến dành cho các mạng không dây thế hệ tiếp theo, sử dụng kiến trúc hai đường dẫn. Mô hình đổi mới này được thiết kế để phân đoạn chính xác các tín hiệu 5G NR và LTE bằng cách xác định nội dung phổ dựa trên tần số và thời gian mà các tín hiệu chiếm dụng. Phương pháp này tích hợp một đường dẫn ngữ cảnh nhằm thu thập thông tin ngữ nghĩa ở mức độ cao, một đường dẫn không gian để bảo toàn các đặc trưng chi tiết về không gian, và một cơ chế hợp nhất đặc trưng mới nhằm kết hợp hiệu quả thông tin từ cả hai đường dẫn. Kiến trúc này có khả năng học tập cả các đặc trưng phổ cục bộ và toàn cục, qua đó nâng cao đáng kể hiệu suất phân đoạn. Kết quả thực nghiệm cho thấy phương pháp này đạt hiệu quả và hiệu suất vượt trội, với một kiến trúc gọn nhẹ chỉ gồm 7 triệu tham số, đạt được độ chính xác toàn cục (global accuracy) là 97.25% và giá trị trung bình của chỉ số giao nhau trên hợp (mean intersection over union – IoU) là 94.76%. Những kết quả này chứng minh rằng đây là một giải pháp đầy hứa hẹn dành cho các hệ thống thông tin không dây thế hệ tiếp

theo.

Mục Lục

1 TỔNG QUAN **1**

1.1 GIỚI THIỆU 1

1.2 MỤC TIÊU 2

1.3 PHƯƠNG PHÁP NGHIÊN CỨU 3

1.4 GIỚI HẠN NGHIÊN CỨU 5

1.5 BỐ CỤC 7

2 CƠ SỞ LÝ THUYẾT **8**

2.1 TỔNG QUAN VỀ GIAO THÔNG VÀ MÔ HÌNH GIAO THÔNG 8

2.1.1 Các tiêu chí cơ bản trong phân tích lưu lượng 8

2.1.2 Các mô hình giao thông 8

2.1.3 Vai trò mô phỏng giao thông trong quản lý đô thị 8

2.2 GIỚI THIỆU XỬ LÝ ẢNH 8

2.3 Giới THIỆU VỀ SUPER RESOLUTION GENERATIVE ADVERSARIAL NETWORK 12

2.3.1 Tổng quan kiến trúc 13

2.3.2 Kiến trúc Generator 13

2.3.3 Kiến trúc Discriminator 14

2.3.4 Thiết kế hàm mất mát 15

2.4 GIỚI THIỆU VỀ ESRGAN VÀ REAL-ESRGAN 17

2.4.1 ESRGAN 17

2.4.2 Real-ESRGAN 18

2.5 GIỚI THIỆU NHẬN DIỆN ĐỐI TƯỢNG 19

2.6 GIỚI THIỆU MÔ HÌNH YOLO 21

2.6.1 Nguyên lý hoạt động của YOLO 22

2.6.2 Hàm mất mát (Loss Function) trong YOLO 24

2.6.3 Các chỉ số đánh giá hiệu suất của YOLO 27

2.6.4 NON-MAXIMUM SUPPRESSION 28

2.7 TỔNG QUAN VỀ MÔ HÌNH YOLOV11 29

2.7.1 Kiến trúc mô hình YOLOv11 30

2.7.2 Tính năng nổi bật của mô hình YOLOv11 31

2.8 Công cụ mô phỏng giao thông: SUMO (Simulation of Urban Mobility) . . 33

2.8.1 Giới thiệu SUMO 33

2.8.2 Ưu điểm và hạn chế của SUMO trong nghiên cứu mô phỏng và thực tế giao thông đô thị 33

2.9 TỔNG QUAN VỀ MÔ HÌNH LSTM 33

3 THIẾT KẾ HỆ THỐNG **34**

3.1 YÊU CẦU CỦA HỆ THỐNG 34

3.2 KIẾN TRÚC HỆ THỐNG 36

3.2.1 Sơ đồ khối hệ thống 36

3.3 Upscale ảnh bằng Real ESRGAN 38

TÀI LIỆU THAM KHẢO **39**

Danh sách hình

2.2.1 Các thành phần cơ bản của xử lý ảnh 11

2.3.2 Kiến trúc SRGAN 13

2.3.3 Kiến trúc của bộ Generator 14

2.3.4 Kiến trúc của bộ Discriminator 15

2.5.5 Nhận diện đối tượng 20

2.6.6 Sơ đồ kiến trúc mạng YOLO 22

2.6.7 Cơ chế hoạt động của YOLO 24

2.6.8 Cách tính chỉ số IOU 28

2.6.9 Kết quả sau khi áp dụng Non-Maximum Suppression 29

2.7.10 Sơ đồ kiến trúc mạng YOLOv11 30

2.7.11 Hiệu suất của mô hình yolov11 so với các phiên bản trước 33

3.2.1 Sơ đồ khối kiến trúc hệ thống 36

Danh sách bảng

Danh sách các từ viết tắt

Các từ viết tắt
None

Chương 1

TỔNG QUAN

1.1 GIỚI THIỆU

Trong bối cảnh đô thị hóa nhanh chóng và sự gia tăng mật độ phương tiện giao thông, việc quản lý và điều tiết giao thông hiệu quả đã trở thành một trong những thách thức lớn nhất mà các thành phố hiện đại phải đối mặt. Theo báo cáo của Tổ chức Hợp tác và Phát triển Kinh tế (OECD), tắc nghẽn giao thông không chỉ gây ra thiệt hại kinh tế hàng tỷ USD mỗi năm mà còn là nguyên nhân chính dẫn đến ô nhiễm không khí, tiêu thụ nhiên liệu không hiệu quả và giảm chất lượng cuộc sống của người dân [1].

Sự phát triển của công nghệ thông tin và truyền thông (ICT) cùng với sự xuất hiện của các hệ thống giao thông thông minh (Intelligent Transportation Systems - ITS) đã mở ra những cơ hội mới trong việc giải quyết các vấn đề giao thông. Đặc biệt, việc ứng dụng các kỹ thuật học máy và học sâu (Deep Learning) vào phân tích dữ liệu giao thông đã cho thấy những kết quả đầy hứa hẹn trong việc dự báo và tối ưu hóa luồng giao thông [2].

Hệ thống camera giao thông hiện đại có khả năng thu thập một lượng lớn dữ liệu hình ảnh theo thời gian thực, tạo ra những "snapshot" phản ánh tình trạng giao thông tại các điểm quan sát. Những dữ liệu này, khi được xử lý bằng các thuật toán học sâu tiên tiến, có thể cung cấp thông tin quý giá về mật độ phương tiện, tốc độ di chuyển, và các mẫu hành vi giao thông, từ đó làm cơ sở cho việc dự báo và điều tiết giao thông hiệu quả.

1.2 MỤC TIÊU

Mục tiêu của đề tài là xây dựng một hệ thống tích hợp chặt chẽ giữa thu thập dữ liệu ảnh từ camera giao thông, xử lý bằng các phương pháp học sâu, và mô phỏng trên nền tảng SUMO để hỗ trợ dự báo tình trạng giao thông và đề xuất các biện pháp điều tiết thông minh. Cụ thể, đề tài nhằm:

- Phát triển giải pháp xử lý ảnh snapshot từ camera giao thông thành các thông số vận tải như lưu lượng xe, tốc độ trung bình, mật độ giao thông và trạng thái các làn đường thông qua mô hình học sâu — từ đó chuyển đổi dữ liệu hình ảnh thô thành dạng đầu vào có cấu trúc cho mô phỏng.
- Tích hợp dữ liệu thu được vào mô hình mô phỏng giao thông trên SUMO để tái hiện tình trạng giao thông thực tế ở khu vực nghiên cứu, từ đó xây dựng một môi trường "in silico" cho phân tích — cho phép đánh giá các kịch bản điều tiết (ví dụ: thay đổi phân bố tín hiệu đèn, điều hướng xe, ưu tiên làn) trước khi áp dụng ngoài thực tế.
- Xây dựng mô hình dự báo ngắn hạn cho tình trạng giao thông (ví dụ: tình trạng ùn tắc, thời gian lưu thông, mật độ xe) dựa trên chuỗi dữ liệu thời gian từ camera và kết quả mô phỏng — nhằm giúp cơ quan quản lý giao thông có khả năng chủ động hơn trong việc ra quyết định.
- Thiết kế và thử nghiệm chiến lược điều tiết giao thông thông minh (ví dụ: điều chỉnh tín hiệu, tái phân làn, điều hướng xe) dựa trên kết quả mô phỏng và dự báo, với mục tiêu giảm thiểu thời gian chờ, mật độ ùn tắc và nâng cao hiệu suất lưu thông tổng thể.
- Đánh giá hiệu quả của hệ thống tích hợp thông qua các thước đo như thời gian lưu thông trung bình, mật độ xe, số lần dừng/chờ, độ chính xác dự báo và khả năng ứng dụng trong thực tế — từ đó đề xuất kiến nghị cho ứng dụng thực tiễn tại đô thị và khả năng mở rộng hệ thống.
- Nâng cao tính linh hoạt và khả năng thích ứng của hệ thống khi phải đối mặt với các điều kiện giao thông thay đổi (giờ cao điểm, sự cố giao thông, điều kiện thời

tiết...) bằng cách tận dụng khả năng học và mô phỏng để điều chỉnh chiến lược điều tiết phù hợp.

Tóm lại, mục tiêu của đề tài là tạo ra một công cụ hỗ trợ thông minh cho quản lý giao thông, kết hợp giữa học sâu và mô phỏng giao thông, giúp từ dữ liệu hình ảnh trực quan chuyển hóa thành các quyết định điều tiết sáng suốt, từ đó cải thiện lưu thông, giảm ùn tắc và nâng cao chất lượng dịch vụ giao thông đô thị.

1.3 PHƯƠNG PHÁP NGHIÊN CỨU

Để đạt được các mục tiêu đã đề ra, Nghiên cứu này áp dụng một quy trình nghiên cứu khoa học chặt chẽ, kết hợp giữa nghiên cứu lý thuyết, thực nghiệm mô phỏng và phân tích đánh giá định lượng. Các phương pháp cụ thể được triển khai như sau:

Phương pháp nghiên cứu lý thuyết:

Trước tiên, sẽ tiến hành khảo sát, tổng hợp và phân tích các kiến thức, lý thuyết nền tảng liên quan tới hệ thống giao thông đô thị, mô hình học sâu (deep learning) cho nhận dạng và đếm phương tiện, mô hình dự báo chuỗi thời gian như Long Short - Term Memory (LSTM), cũng như mô phỏng giao thông (traffic simulation) với SUMO. Cụ thể, sẽ tổng hợp các nghiên cứu về việc sử dụng mô hình thí dụ như YOLOv8 để phát hiện và đếm phương tiện từ camera giao thông (ví dụ các nghiên cứu cho thấy YOLOv8 được ứng dụng trong thực tế để đếm xe và tính mật độ giao thông). Đồng thời, tìm hiểu các phương pháp mô phỏng kết hợp camera/nhận dạng + mô hình điều tiết tín hiệu giao thông, các nghiên cứu tích hợp giữa phát hiện hình ảnh và mô phỏng như kết hợp giữa camera và SUMO. Qua đó, xây dựng cơ sở lý thuyết cho việc lựa chọn kiến trúc hệ thống, xác định các tham số chính, và xác định các chỉ tiêu đánh giá hiệu quả (thời gian lưu thông, mật độ, số lần dừng/chờ, độ chính xác dự báo, v.v.).

Phương pháp thực nghiệm mô phỏng:

- **Thu thập dữ liệu snapshot camera và xử lý bằng học sâu:** Sử dụng mô hình YOLOv8 để phát hiện và đếm phương tiện từ ảnh hoặc video snapshot thu được

từ camera giao thông. Từ kết quả đếm được xác định các thông số như lưu lượng xe, mật độ theo từng mốc thời gian và từng điểm giao thông. Đây chính là bước chuyển dữ liệu hình ảnh thô thành dạng dữ liệu có cấu trúc để sử dụng tiếp. Việc này tận dụng các thư viện, công cụ thực nghiệm đã có (ví dụ từ thực tế/nguồn mở) như một số nghiên cứu đã thực hiện.

- **Xây dựng mô hình dự báo mật độ giao thông:** Từ chuỗi dữ liệu mật độ theo mốc thời gian thu được, tiến hành huấn luyện mô hình LSTM để dự báo mật độ giao thông cho 15 phút tiếp theo. Việc này gồm tiền xử lý dữ liệu (chuỗi thời gian, tạo các đặc trưng như thời gian, ngày, giờ, điểm giao thông, loại phương tiện nếu có), chia train/validation/test, lựa chọn cấu trúc mạng LSTM (số lớp, số đơn vị, dropout, epochs...), và kiểm định mô hình qua các chỉ tiêu như RMSE, MAE, MAPE.

- **Tích hợp mô phỏng giao thông với SUMO:** Sử dụng môi trường SUMO để tái hiện mạng lưới giao thông nghiên cứu, cấu hình các tham số như làn đường, tín hiệu giao thông, các điểm camera/tuyến đường tương ứng với dữ liệu thực tế. Sau đó, nhập dữ liệu mật độ xe thực tế (hoặc dữ liệu từ mô hình đếm) để khớp mô phỏng sao cho trạng thái mô phỏng càng sát thực càng tốt. Đây là bước "đồng bộ" giữa dữ liệu thực và mô phỏng.

- **Tích hợp module điều khiển giao thông:** Xây dựng thuật toán điều tiết giao thông (ví dụ: điều chỉnh tín hiệu, phân luồng, ưu tiên làn) dựa trên dự báo mật độ từ LSTM và kết quả mô phỏng. Thử chạy hai kịch bản: kịch bản không điều tiết (tín hiệu cố định hoặc theo chế độ hiện hữu) và kịch bản có điều tiết (tín hiệu và phân luồng thay đổi dựa trên dự báo và mô phỏng). Chạy mô phỏng SUMO cho cả hai kịch bản và thu thập dữ liệu kết quả.

- **Xây dựng dashboard trực quan:** Triển khai giao diện hiển thị gồm luồng xe chạy như mô phỏng (trong SUMO), biểu đồ mật độ theo thời gian, kết quả dự báo LSTM, và so sánh hiệu quả giữa hai kịch bản (có/không điều tiết). Dashboard sẽ giúp trực quan hóa kết quả và hỗ trợ phân tích.

- **Rút ra kết luận, đề xuất cải tiến và kiến nghị ứng dụng thực tiễn:** từ kết quả thực nghiệm và mô phỏng, chỉ rõ giới hạn của nghiên cứu, gợi ý mở rộng (ví dụ

mở rộng mạng lưới, loại phương tiện đa dạng, tích hợp dữ liệu thời tiết, sự cố...).

Phương pháp phân tích đánh giá:

- **Phân tích độ chính xác của bước phát hiện và đếm phương tiện từ camera:** So sánh với đếm thủ công hoặc dữ liệu tham chiếu nếu có.
- **Phân tích hiệu năng dự báo của mô hình LSTM:** Sử dụng các chỉ tiêu như RMSE, MAE, MAPE, R^2... để đánh giá khả năng dự báo mật độ 15 phút tới.
- **Phân tích kết quả mô phỏng SUMO:** So sánh giữa kịch bản không điều tiết và có điều tiết trên các chỉ tiêu như thời gian lưu thông trung bình, mật độ xe, số lần dừng/chờ, độ ổn định luồng giao thông.
- **Đánh giá tổng thể hệ thống:** Xem xét khả năng tích hợp giữa các thành phần (đếm - dự báo - mô phỏng - điều tiết), tính khả thi triển khai thực tế, độ linh hoạt khi điều kiện giao thông thay đổi (giờ cao điểm, sự cố...).
- **Trực quan hóa kết quả trên dashboard:** So sánh biểu đồ, luồng xe, và rút ra nhận xét về hiệu quả điều tiết, lợi ích đối với quản lý giao thông.

1.4 GIỚI HẠN NGHIÊN CỨU

Hệ thống nghiên cứu này mặc dù cố gắng tích hợp các thành phần thu thập dữ liệu hình ảnh, học sâu, dự báo và mô phỏng điều tiết giao thông, nhưng vẫn tồn tại một số giới hạn đáng lưu ý. Trước hết, việc sử dụng ảnh snapshot từ camera giao thông chỉ phản ánh trạng thái "tĩnh" tại các thời điểm chụp và phụ thuộc vào điều kiện quan sát như góc đặt camera, ánh sáng, che khuất, chất lượng hình ảnh. Do đó, khả năng phát hiện và đếm phương tiện bằng mô hình YOLOv8 có thể bị ảnh hưởng bởi các biến môi trường (như mưa, sương, bóng đổ) hoặc khu vực giao thông phức tạp (nhiều phương tiện, làn không phân định rõ). Tiếp theo, dữ liệu thu được tại một hoặc một số điểm giao thông cụ thể có thể không đại diện cho toàn bộ mạng lưới giao thông hoặc các điều kiện giao thông khác nhau về giờ cao điểm, ngày lễ, sự cố - dẫn đến hạn chế trong khái quát hóa kết quả.

Về phần mô hình học sâu và dự báo bằng Long Short-Term Memory (LSTM), mặc dù có khả năng dự báo mật độ giao thông cho 15 phút tiếp theo, nhưng việc huấn luyện chỉ trên dữ liệu từ các mốc thời gian cố định và điểm giao thông nhất định khiến mô hình có thể kém chính xác khi đối mặt với tình huống bất thường (như tai nạn, thay đổi bất ngờ lưu lượng, thời tiết cực đoan) mà không có dữ liệu học trước. Hơn nữa, mô hình dự báo chỉ tập trung vào một biến chính - mật độ phương tiện - và chưa tích hợp đầy đủ các yếu tố khác như loại phương tiện, tốc độ, hành vi người lái, ảnh hưởng từ tín hiệu đèn hoặc thay đổi hành lang giao thông.

Phần mô phỏng giao thông với SUMO dù được "khớp" với dữ liệu thực tế đến mức có thể nhưng vẫn có giới hạn vì mô phỏng luôn là bản sao không hoàn hảo của môi trường thực. Mạng lưới mô phỏng có thể chưa mô hình hóa đầy đủ mọi chiều của thực tế như sự tương tác phức tạp giữa các phương tiện, hành vi bất định, ảnh hưởng thời tiết, người đi bộ, xe máy nhỏ, hoặc việc vi phạm giao thông - những yếu tố rất phổ biến tại đô thị như Hồ Chí Minh. Việc điều tiết giao thông thông qua thuật toán cũng đặt giả định rằng các thông số mô hình và dữ liệu đầu vào là đồng nhất và ổn định, trong khi thực tế có thể thay đổi nhanh và không thể đo trước hết.

Cuối cùng, việc xây dựng dashboard trực quan để hiển thị kết quả mô phỏng, dự báo và điều tiết cũng gặp giới hạn do khả năng phản ánh toàn bộ thực tế - dashboard chỉ hiển thị dữ liệu và mô phỏng ở mức độ "có thể" và phù hợp với giả định nghiên cứu. Những quyết định điều tiết đưa ra từ mô hình có thể chưa tính tới đầy đủ chi phí thực thi, điều kiện vận hành thực tế, phản ứng của người tham giao thông hoặc các yếu tố tổ chức giao thông ngoài mô hình.

Tóm lại, các giới hạn nghiên cứu chính bao gồm: dữ liệu nguồn (chỉ snapshot camera, tại các điểm giới hạn), khả năng khái quát hóa kết quả mô hình và mô phỏng, độ chính xác và phạm vi của mô hình dự báo, sự đơn giản hóa môi trường mô phỏng và giả định điều tiết, và khả năng ứng dụng thực tế bị chi phối bởi nhiều yếu tố ngoài mô hình. Hiểu và thừa nhận các giới hạn này giúp bạn đọc đánh giá đúng mức độ đóng góp của nghiên cứu, và tạo nền tảng cho các nghiên cứu tiếp theo.

1.5 BỐ CỤC

Chương 1: Tổng quan - Trình bày bối cảnh, động lực nghiên cứu, mục tiêu và phương pháp nghiên cứu.

Chương 2: Cơ sở lý thuyết - Tổng quan các kiến thức nền tảng về xử lý ảnh, học sâu, dự báo chuỗi thời gian và học tăng cường.

Chương 3: Thiết kế hệ thống - Trình bày kiến trúc tổng thể của hệ thống và thiết kế chi tiết các mô-đun.

Chương 4: Kết quả vào thảo luận - Mô tả quá trình triển khai hệ thống và các thí nghiệm đánh giá.

Chương 5: Kết luận và hướng phát triển - Tổng kết những đóng góp của nghiên cứu và đề xuất hướng phát triển tương lai.

Chương 2

CƠ SỞ LÝ THUYẾT

2.1 TỔNG QUAN VỀ GIAO THÔNG VÀ MÔ HÌNH GIAO THÔNG

2.1.1 Các tiêu chí cơ bản trong phân tích lưu lượng

2.1.2 Các mô hình giao thông

2.1.3 Vai trò mô phỏng giao thông trong quản lý đô thị

2.2 GIỚI THIỆU XỬ LÝ ẢNH

Xử lý ảnh (Image Processing) là lĩnh vực thuộc khoa học máy tính và kỹ thuật, tập trung vào việc phân tích và biến đổi hình ảnh nhằm cải thiện chất lượng hoặc trích xuất thông tin phục vụ quan sát và nhận dạng. Đây là nền tảng quan trọng của thị giác máy tính, vì hầu hết các thuật toán phân tích hay học máy đều yêu cầu dữ liệu hình ảnh đã được xử lý chuẩn hóa. [3]

Về bản chất, xử lý ảnh là quá trình thao tác trực tiếp trên ma trận điểm ảnh để làm rõ thông tin hữu ích và loại bỏ nhiễu hay các thành phần không cần thiết. Các kỹ thuật này thường được chia làm hai mức: xử lý ảnh mức thấp và mức cao. Mức thấp chủ yếu gồm các phép biến đổi tín hiệu như lọc nhiễu, điều chỉnh độ sáng, thay đổi không

gian màu hoặc tăng độ sắc nét mà không xét đến nội dung ảnh. Trong khi đó, xử lý ảnh mức cao tập trung vào việc trích xuất đặc trưng như biên cạnh, điểm đặc trưng, kết cấu hay hình dạng, tạo dữ liệu có ý nghĩa cho các mô hình nhận diện và học sâu. Các thành phần cơ bản của xử lý ảnh:

- **Hệ thống thu nhận hình ảnh (Image Sensor):** Để thu được ảnh số, cần hai thành phần cơ bản. Thứ nhất là cảm biến vật lý, có khả năng phản hồi với năng lượng phát ra từ đối tượng quan sát. Thứ hai là thiết bị chuyển đổi tín hiệu từ cảm biến sang dạng số, thường gọi là bộ số hóa (digitizer). Bộ số hóa này chịu trách nhiệm biến các tín hiệu thu được từ cảm biến thành dữ liệu số để máy tính có thể xử lý.

- **Phần cứng xử lý ảnh chuyên dụng (Specialized Image Processing Hardware):** Để thực hiện các phép tính số học và logic trên toàn bộ ảnh, cần kết hợp giữa bộ số hóa và phần cứng chuyên dụng, thường được gọi là hệ thống tiền xử lý (front-end subsystem). Tốc độ xử lý của phần cứng này là yếu tố quan trọng nhất, vì các máy tính thông thường khó đáp ứng được yêu cầu truyền dữ liệu với tốc độ cao cần thiết cho xử lý ảnh thời gian thực.

- **Máy tính: (Computer):** Máy tính trong hệ thống xử lý ảnh là máy tính đa năng, có thể là một PC thông thường hoặc siêu máy tính tùy vào quy mô ứng dụng. Một máy tính cá nhân cấu hình tốt thường đủ cho các tác vụ xử lý ảnh offline, phục vụ nghiên cứu và phân tích dữ liệu hình ảnh.

- **Phần mềm xử lý ảnh (Image Processing Software):** Phần mềm xử lý ảnh gồm các module chuyên dụng thực hiện những nhiệm vụ cụ thể. Một bộ phần mềm thiết kế tốt sẽ cho phép người dùng viết ít lệnh nhất, đồng thời tận dụng tối đa các module có sẵn. Các gói phần mềm phát triển cao còn cho phép tích hợp các module và câu lệnh lập trình từ ít nhất một ngôn ngữ lập trình. Ví dụ, MATLAB là một trong những công cụ phổ biến được dùng trong các hệ thống xử lý ảnh.

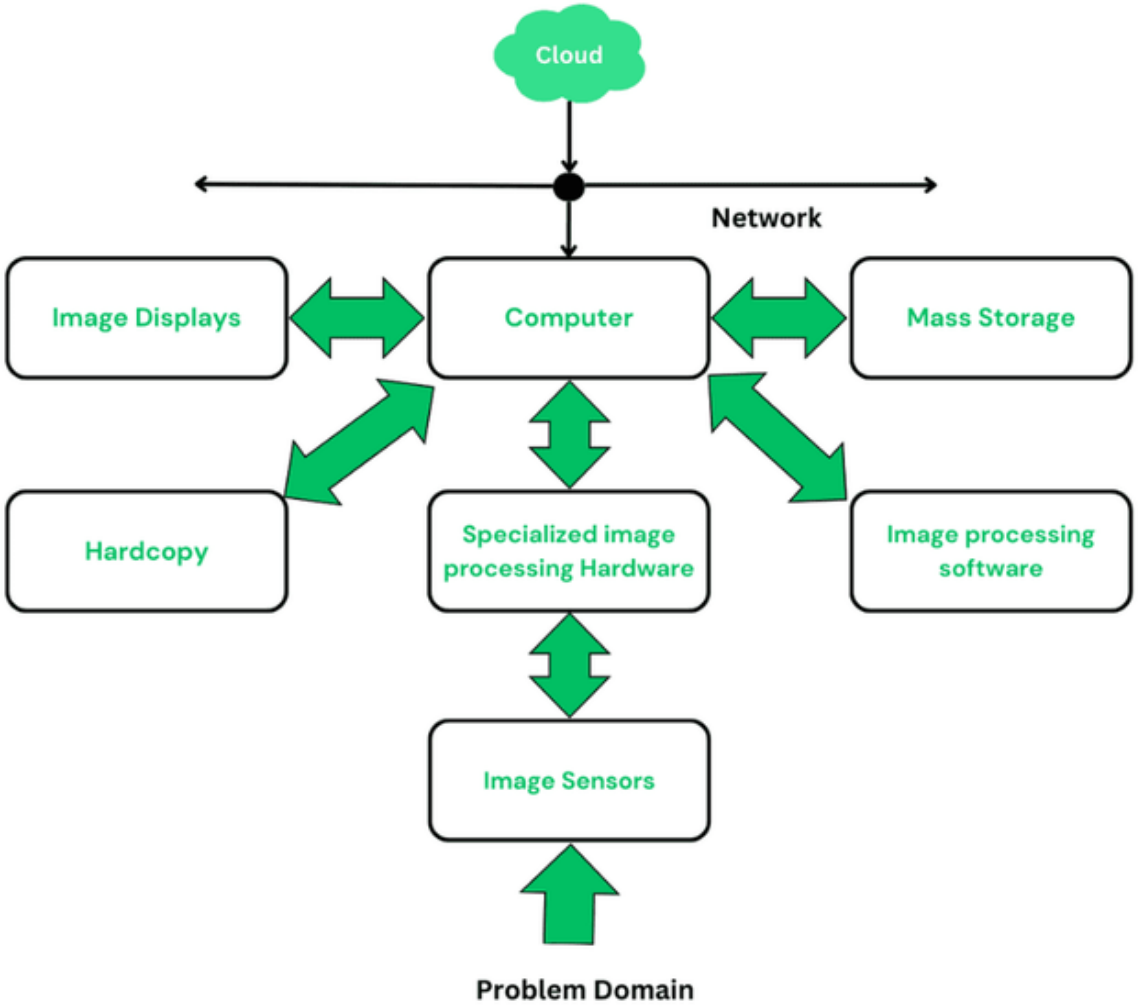
- **Lưu trữ dữ liệu (Mass Storage):** Lưu trữ là yếu tố quan trọng trong xử lý ảnh, đặc biệt khi làm việc với các ảnh có dung lượng lớn. Ví dụ, một ảnh có kích thước 1024 x 1024 pixel cần khoảng 1 megabyte nếu chưa nén. Các hệ thống xử lý ảnh thường phải lưu trữ hàng nghìn hoặc thậm chí hàng triệu ảnh. Hệ thống lưu trữ số

tuân theo ba nguyên tắc cơ bản: Lưu trữ tạm thời (dùng trong quá trình xử lý), lưu trữ trực tuyến (phục vụ truy suất nhanh) và lưu trữ lâu dài (truy xuất ít, lưu trữ dài hạn.).

- **Hiển thị hình ảnh (Image Displays):** Màn hình hiển thị thường là màn hình màu phẳng, được điều khiển bởi card đồ họa hoặc card hiển thị hình ảnh. Đây là thành phần quan trọng giúp máy tính trình chiếu và thao tác với dữ liệu hình ảnh.

- **Thiết bị in ấn (Hardcopy):** Để lưu trữ hoặc trình bày hình ảnh, có thể dùng các thiết bị in ấn và ghi hình, bao gồm máy in laser, máy ảnh phim, thiết bị nhạy nhiệt, máy in phun, hoặc các phương tiện số như ổ đĩa quang và CD-ROM. Phim ảnh cung cấp độ phân giải cao nhất, trong khi giấy là phương tiện dễ sử dụng để trình bày nội dung. Khi dùng thiết bị chiếu ảnh số, hình ảnh vẫn tồn tại dưới dạng dữ liệu số, giúp dễ dàng trình chiếu hoặc lưu trữ lâu dài.

- **Mạng và điện toán đám mây (Cloud):** Trong thời đại hiện nay, mạng và điện toán đám mây là những yếu tố thiết yếu trong xử lý ảnh. Vì dữ liệu hình ảnh thường có dung lượng rất lớn, băng thông trở thành vấn đề quan trọng khi truyền tải. Khi gửi dữ liệu qua Internet đến các địa điểm từ xa, hiệu quả truyền tải không phải lúc nào cũng cao, do đó công nghệ cáp quang và các giải pháp băng thông rộng được sử dụng. Bên cạnh đó, nén dữ liệu ảnh đóng vai trò quan trọng để giảm dung lượng truyền tải, giúp gửi lượng lớn hình ảnh một cách nhanh chóng và hiệu quả. [4]



Hình 2.2.1: Các thành phần cơ bản của xử lý ảnh

Trong bối cảnh ứng dụng hiện đại, đặc biệt là các hệ thống thông minh như theo dõi giao thông, giám sát đô thị hay phân tích dữ liệu từ camera, xử lý ảnh giữ vai trò như một giai đoạn tiền xử lý không thể thiếu. Ví dụ, dữ liệu từ camera giao thông thường gặp nhiều hạn chế: ánh sáng thay đổi liên tục, điều kiện thời tiết gây nhiễu, độ phân giải không đồng đều, và vật thể thường nhỏ hoặc bị che khuất. Nếu không có bước xử lý ảnh phù hợp, những hạn chế này sẽ ảnh hưởng trực tiếp đến độ chính xác của các mô hình phát hiện và đếm phương tiện. Các kỹ thuật phổ biến như cân bằng histogram, lọc Gaussian, chuyển đổi sang ảnh xám, khử nhiễu bằng median filter hay tăng độ tương phản đều được áp dụng để cải thiện độ rõ ràng trước khi đưa ảnh vào pipeline phân tích.

Sự phát triển của học sâu trong hơn một thập kỷ qua đã mở ra một hướng mới cho xử lý ảnh, nơi mà các mô hình không chỉ thực hiện các phép biến đổi dựa trên quy

tắc cố định mà còn học trực tiếp từ dữ liệu để tối ưu hóa chất lượng đầu ra. Các bài toán như khử nhiễu (denoising), tăng độ phân giải (super-resolution), tái tạo ảnh thiếu thông tin (inpainting), tách nền và nhiều dạng biến đổi phức tạp khác đều đạt được chất lượng vượt trội nhờ mạng nơ-ron tích chập (CNN) và các mô hình sinh ảnh như GAN. Đặc biệt, các hệ thống giám sát giao thông sử dụng camera có thể hưởng lợi trực tiếp từ những kỹ thuật này: ảnh từ camera độ phân giải thấp có thể được tăng cường bằng super-resolution, giúp mô hình phát hiện phương tiện hoạt động chính xác hơn trong môi trường phức tạp. [5]

Ngoài ra, xử lý ảnh còn liên quan chặt chẽ đến việc chuẩn hóa dữ liệu đầu vào cho các thuật toán học máy. Việc thay đổi kích thước, chuẩn hóa pixel theo phân phối thống nhất, hoặc điều chỉnh tỷ lệ khung hình đều giúp giảm tải tính toán và tăng độ ổn định khi huấn luyện mô hình nhận dạng hoặc dự đoán. Điều này đặc biệt quan trọng trong các hệ thống thời gian thực, nơi tốc độ xử lý và độ ổn định đóng vai trò quyết định.

Tóm lại, xử lý ảnh là bước khởi đầu cho mọi đề tài thị giác máy tính, cung cấp nền tảng và dữ liệu chất lượng cho các thuật toán học sâu, phát hiện đối tượng và phân tích hành vi. Với sự phát triển nhanh chóng của các mô hình hiện đại, xử lý ảnh không chỉ còn dừng lại ở các kỹ thuật truyền thống mà đang chuyển mình mạnh mẽ và đóng vai trò quan trọng trong việc xây dựng những hệ thống thông minh, chính xác và tin cậy — đặc biệt trong lĩnh vực mô phỏng và quản lý giao thông dựa trên dữ liệu từ camera.

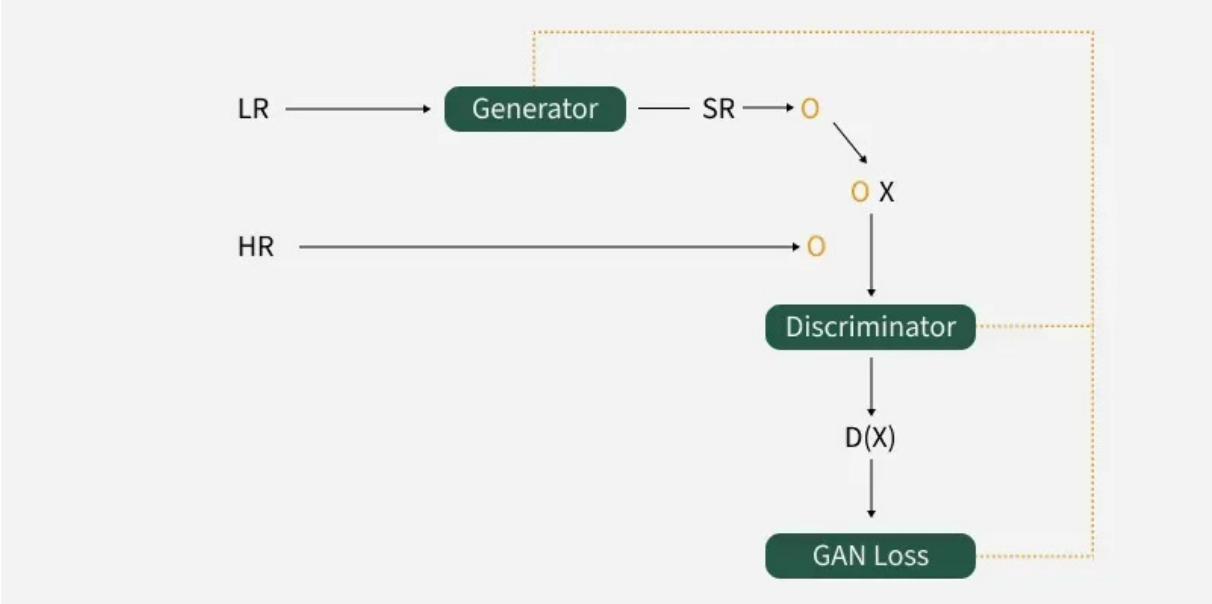
2.3 Giới THIỆU VỀ SUPER RESOLUTION GENERATIVE ADVERSARIAL NETWORK

Super-Resolution Generative Adversarial Network (SRGAN), được Ledig và cộng sự giới thiệu vào năm 2016, là một trong những mô hình tiên phong ứng dụng mạng GAN vào bài toán tăng độ phân giải ảnh. Mục tiêu của SRGAN là khắc phục hạn chế của các phương pháp nội suy truyền thống và các mô hình tối ưu theo lỗi điểm ảnh như MSE, vốn thường tạo ra ảnh mượt nhưng thiếu chi tiết và không giữ được kết cấu tự nhiên. Thông qua cơ chế huấn luyện đối kháng, SRGAN học cách sinh ra ảnh có độ phân giải cao với chi tiết sắc nét hơn bằng cách sử dụng đồng thời hai thành phần: perceptual loss dựa trên đặc trưng trích xuất từ mạng VGG và adversarial loss từ Discriminator. Sự kết

hợp này giúp mô hình tái tạo các hoa văn, kết cấu và đường nét tinh vi thường bị mất đi trong quá trình phóng to ảnh, từ đó tạo ra ảnh đầu ra có chất lượng thị giác chân thực và giàu chi tiết hơn so với các kỹ thuật SR truyền thống. [6]

2.3.1 Tổng quan kiến trúc

SRGAN hoạt động theo cơ chế GAN truyền thống, trong đó có hai mạng nơ-ron đóng vai trò đối kháng nhau: Generator nhận ảnh độ phân giải thấp và sinh ra phiên bản độ phân giải cao, trong khi Discriminator cố gắng phân biệt ảnh thật với ảnh được tạo ra. Quá trình huấn luyện xen kẽ này buộc Generator phải liên tục cải thiện chất lượng ảnh sinh ra để ngày càng giống với ảnh thật hơn. [7]



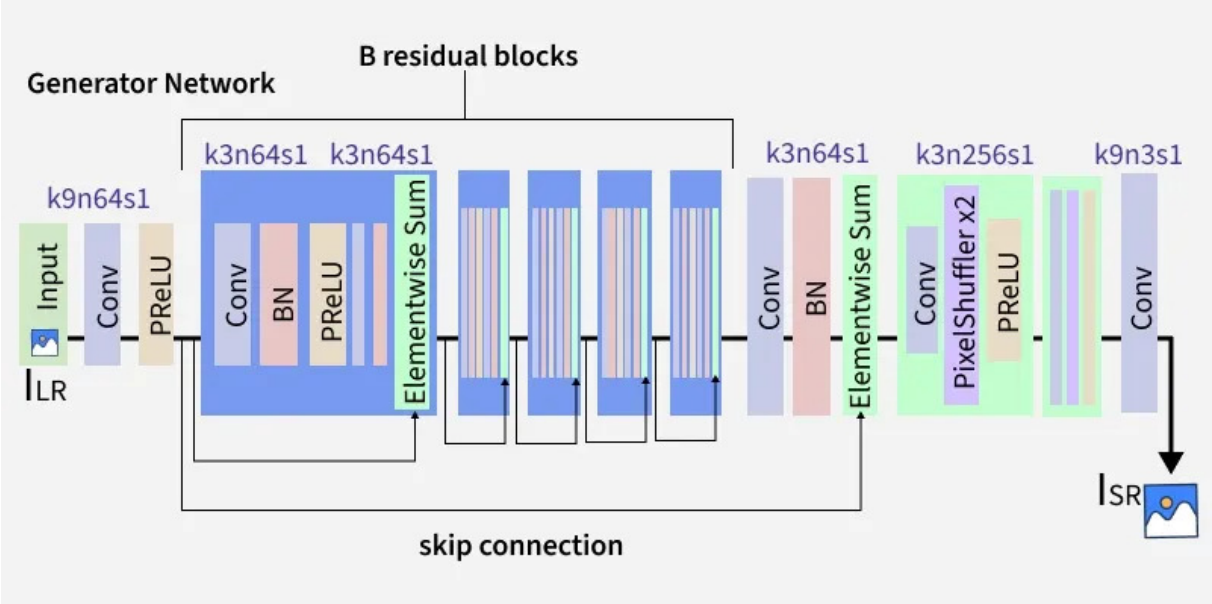
Hình 2.3.2: Kiến trúc SRGAN

2.3.2 Kiến trúc Generator

Generator sử dụng kiến trúc mạng Residual Network (ResNet) thay vì các mạng tích chập sâu thông thường. Việc lựa chọn ResNet đóng vai trò quan trọng bởi các kết nối tắt (skip connections) trong kiến trúc này giúp dòng gradient lan truyền hiệu quả hơn trong quá trình huấn luyện. Nhờ đó, mô hình có thể xây dựng các mạng rất sâu mà không gặp phải hiện tượng mất mát gradient, đồng thời cải thiện khả năng học đặc trưng tinh vi trong ảnh.

Bộ sinh (Generator) được xây dựng từ 16 Residual Block, mỗi khối gồm hai lớp tích chập với kernel kích thước 3x3 và 64 kênh đặc trưng. Sau mỗi lớp tích chập là Batch Normalization và hàm kích hoạt Parametric ReLU (PReLU). Khác với ReLU hoặc LeakyReLU truyền thống, PReLU cho phép hệ số dốc ở vùng âm được học tự động, giúp mô hình thích nghi tốt hơn trong quá trình huấn luyện mà không làm tăng nhiều chi phí tính toán.

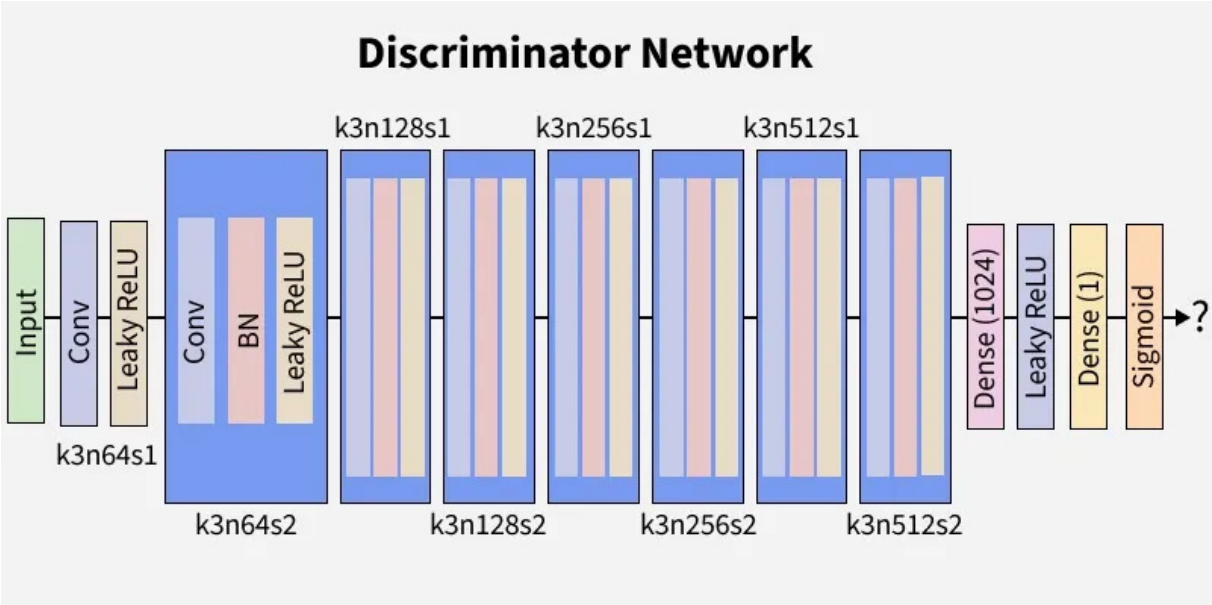
Giai đoạn tăng độ phân giải của mô hình được thực hiện thông qua hai lớp sub-pixel convolution đã được huấn luyện, giúp phóng to kích thước không gian một cách hiệu quả. Phương pháp này hoạt động bằng cách tái sắp xếp thông tin từ chiều kênh sang chiều không gian, cho phép mô hình học trực tiếp cách upsample ảnh thay vì sử dụng các kỹ thuật nội suy đơn giản.



Hình 2.3.3: Kiến trúc của bộ Generator

2.3.3 Kiến trúc Discriminator

Bộ phân biệt (Discriminator) được thiết kế theo kiến trúc sâu gồm tám lớp tích chập 3x3, trong đó số lượng đặc trưng được mở rộng dần từ 64 lên 512 khi kích thước không gian giảm xuống thông qua các lớp tích chập có bước nhảy (strided convolution). Sau khối trích xuất đặc trưng này, mô hình sử dụng hai lớp kết nối đầy đủ (fully connected) và kết thúc bằng hàm kích hoạt sigmoid, nhằm đưa ra xác suất thể hiện mức độ "thật" của ảnh đầu vào—tức phân biệt ảnh thực với ảnh do bộ sinh tạo ra.



Hình 2.3.4: Kiến trúc của bộ Discriminator

2.3.4 Thiết kế hàm mất mát

Content Loss

Trong các phương pháp super-resolution truyền thống, Mean Squared Error (MSE) thường được sử dụng làm hàm mất mát, đo lường sai khác điểm ảnh giữa ảnh sinh và ảnh đích. Tuy nhiên, việc tối ưu theo MSE thường khiến ảnh tái tạo trở nên quá trơn mượt, thiếu chi tiết. Nguyên nhân là MSE thúc đẩy mô hình tạo ra một kết quả "trung bình" trong số nhiều khả năng khôi phục ảnh độ phân giải cao tương ứng với một ảnh đầu vào bị giảm chất lượng, từ đó làm mất đi các cấu trúc tinh tế và độ sắc nét vốn có.

$$l^{SR}_{VGG/i,j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y}\right)^2 \tag{2.3.1}$$

Trong đó:

- $l^{SR}_{VGG/i,j}$: Giá trị hàm mất mát (VGG) tại lớp (i, j).
- $W_{i,j}, H_{i,j}$: Chiều rộng và chiều cao của bản đồ đặc trưng VGG tại tầng (i, j), dùng để chuẩn hóa.
- $\phi_{i,j}$: Bản đồ đặc trưng được trích xuất từ lớp (i, j) của mạng VGG đã được huấn

luyện trước.

- I^{HR}: Ảnh độ phân giải cao thực tế (ground-truth).
- I^{LR}: Ảnh đầu vào có độ phân giải thấp.
- $G_{\theta_G}(I^{LR})$: Ảnh độ phân giải cao được sinh ra bởi bộ Generator G.
- (x, y): Vị trí không gian một điểm trong bản đồ đặc trưng.

Adversarial Loss

Adversarial loss đóng vai trò thúc đẩy bộ sinh tạo ra các ảnh mà bộ phân biệt không thể phân tách với ảnh độ phân giải cao thực sự. Thành phần mất mát này đặc biệt quan trọng trong việc khôi phục các chi tiết sắc nét và kết cấu chân thực, giúp ảnh được phóng đại trở nên tự nhiên và thuyết phục hơn về mặt thị giác.

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{2.3.2}$$

Trong đó:

- l_{Gen}^{SR}: Giá trị Adversarial loss cho bộ Generator.
- N: Số lượng mẫu ảnh.
- $D_{\theta_D}(G_{\theta_G}(I^{LR}))$: Xác suất mà bộ Discriminator đánh giá ảnh được sinh ra là ảnh thật.
- $G_{\theta_G}(I^{LR})$: Ảnh độ phân giải cao được sinh ra bởi bộ Generator sử dụng ảnh đầu vào có độ phân giải thấp I^{LR}.
- $-logD_{\theta D}(G_{\theta G}(I^{LR}))$: Phạt Generator khi Discriminator dễ dàng nhận ra ảnh giả mà nó tạo ra.

Tổng hợp hàm mất mát

$$l^{SR} = l_X^{SR} + 10^{-3} l_{Gen}^{SR} \tag{2.3.3}$$

Trong đó:

- l^{SR}: Hàm mất mát tổng hợp cho super-resolution.
- l_X^{SR}: Content loss dựa trên đặc trưng VGG perceptual.
- l_{Gen}^{SR}: Adversarial loss từ bộ Generator.
- 10^{-3}: Hệ số trọng số cân bằng giữa hai thành phần mất mát, đảm bảo Content Loss chiếm ưu thế trong quá trình huấn luyện.

2.4 GIỚI THIỆU VỀ ESRGAN VÀ REAL-ESRGAN

2.4.1 ESRGAN

ESRGAN (Enhanced Super-Resolution GAN) là phiên bản cải tiến của SRGAN, tiếp tục khai thác sức mạnh của mô hình GAN trong bài toán tăng độ phân giải ảnh. ESRGAN khắc phục những hạn chế của SRGAN, đồng thời nâng cao chất lượng thị giác bằng cách tái tạo nhiều chi tiết tinh vi và sắc nét hơn. Những cải tiến đáng chú ý của ESRGAN gồm: [8]

- **Residual-in-Residual Dense Block (RRDB):** Cấu trúc khối mới giúp tăng cường khả năng học đặc trưng và cải thiện độ ổn định khi huấn luyện, thay thế cho các residual block truyền thống.
- **Nâng cấp perceptual loss:** Giúp mô hình tạo ra hình ảnh tự nhiên và chân thực hơn.
- **Relativistic GAN:** ESRGAN sử dụng hàm mất mát GAN mang tính tương đối (Relativistic GAN loss) thay cho GAN cổ điển, nhằm giúp bộ phân biệt đánh giá ảnh thật có "tính chân thực cao hơn tương đối" so với ảnh giả, thay vì chỉ đánh giá theo dạng nhị phân thật/giả.

Các thành phần chính trong cấu trúc:

- **Generator:** Nhiệm vụ của bộ sinh là chuyển đổi ảnh đầu vào độ phân giải thấp (LR) thành ảnh độ phân giải cao (HR). Trong kiến trúc ESRGAN, các khối RRDB đóng vai trò trung tâm trong việc trích xuất đặc trưng và phóng đại ảnh, đây cũng là điểm cải tiến quan trọng so với SRGAN.

- **Discriminator:** Bộ phân biệt được huấn luyện để phân loại xem một ảnh là ảnh HR thật hay ảnh SR do generator tạo ra. Mục tiêu của nó là phát hiện chính xác các ảnh giả nhằm thúc đẩy generator tạo ra hình ảnh ngày càng chân thực.

- **Perceptual Loss:** Một cải tiến quan trọng của ESRGAN là sử dụng hàm mất mát cảm nhận, đo mức độ tương đồng thị giác giữa ảnh sinh và ảnh thật bằng cách so sánh các bản đồ đặc trưng trích xuất từ mạng VGG đã được huấn luyện trước. Điều này giúp ảnh SR có tính tự nhiên và dễ chịu hơn đối với người quan sát.

Các hàm mất mát trong ESRGAN:

- **Content Loss:** Đo sự khác biệt giữa ảnh HR thật và ảnh HR được mô hình tạo ra, thường tính theo mức pixel với chỉ số MSE.

- **Adversarial Loss:** Bảo đảm rằng ảnh sinh trông càng giống ảnh thật càng tốt bằng việc tối ưu dựa trên phản hồi từ discriminator.

- **Perceptual Loss:** So sánh các đặc trưng bậc cao giữa ảnh sinh và ảnh thật, nhằm duy trì chất lượng thị giác, chi tiết và kết cấu trong ảnh.

2.4.2 Real-ESRGAN

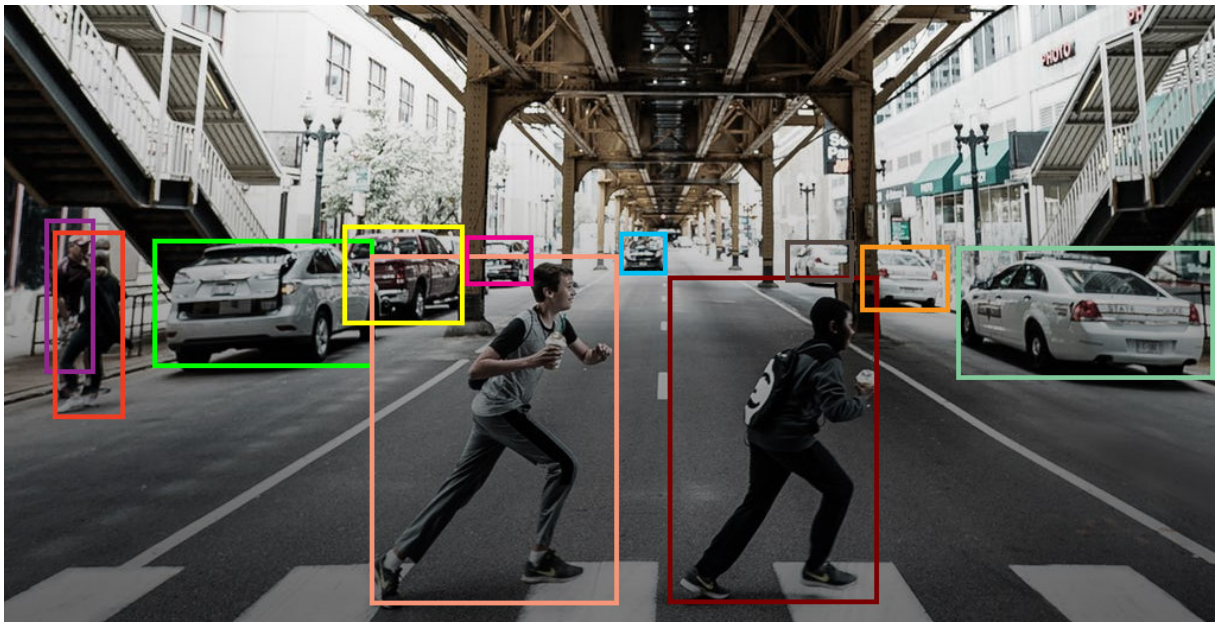
Real-ESRGAN (Real-World Enhanced Super-Resolution Generative Adversarial Network) được phát triển nhằm khắc phục các hạn chế của ESRGAN khi xử lý dữ liệu ngoài đời thực. Trong khi ESRGAN chủ yếu hoạt động tốt trên các tập dữ liệu tổng hợp và giả lập—nơi nhiễu, mờ và suy giảm được mô hình hóa đơn giản—Real-ESRGAN hướng tới việc tái tạo ảnh độ phân giải cao từ các đầu vào bị suy giảm phức tạp và không dự đoán được trong thực tế. Mô hình này sử dụng cơ chế GAN bất đối xứng (asymmetric GAN) với bộ suy giảm mạnh mẽ (degradation model) gồm nhiều giai đoạn, mô phỏng các hiện

tượng suy giảm thường gặp như nhiễu cảm biến, nén JPEG, mất chi tiết do chuyển động và sai lệch quang học.

Real-ESRGAN áp dụng kiến trúc RRDB (Residual-in-Residual Dense Block) tương tự ESRGAN nhưng được điều chỉnh để tăng tính ổn định và độ bền khi huấn luyện trên dữ liệu phi chuẩn hóa. Ngoài ra, Real-ESRGAN giới thiệu bộ phân biệt cải tiến (U-shaped discriminator) giúp mô hình học được nhiều mức độ suy giảm khác nhau. Nhờ các cải tiến này, Real-ESRGAN có khả năng tái tạo chi tiết tốt hơn, ổn định hơn trên ảnh chụp thực tế, và cung cấp kết quả nhất quán ngay cả trong điều kiện ảnh đầu vào bị hỏng nghiêm trọng.

2.5 GIỚI THIỆU NHẬN DIỆN ĐỐI TƯỢNG

Một trong những lĩnh vực trọng tâm của Trí tuệ nhân tạo (Artificial Intelligence) là Thị giác máy tính (Computer Vision). Đây là ngành nghiên cứu các phương pháp thu nhận, xử lý và phân tích ảnh số nhằm mô phỏng khả năng "nhìn" và "hiểu" thế giới trực quan của con người. Computer Vision bao gồm nhiều bài toán quan trọng như phân đoạn ảnh, nhận dạng đối tượng, mô phỏng cảnh, siêu phân giải hình ảnh, tái tạo 3D và nhiều hướng tiếp cận khác. Trong số đó, Object Detection được xem là một trong những bài toán cốt lõi và có tác động lớn nhất nhờ tính ứng dụng rộng rãi trong thực tiễn.



Hình 2.5.5: Nhận diện đối tượng

Một số ứng dụng của nhận diện đối tượng bao gồm:

- Phân tích thể thao: Theo dõi vị trí và chuyển động của cầu thủ trên sân.
- Y tế: Phát hiện bất thường trong hình ảnh y khoa.
- Giao thông thông minh: Nhận diện phương tiện, biển báo giao thông, phát hiện vi phạm.
- Thương mại điện tử: Tìm kiếm sản phẩm theo hình ảnh, kiểm tra hàng hóa tự động.
- Nông nghiệp: Giám sát cây trồng, phát hiện sâu bệnh qua hình ảnh.
- Robot tự hành: Giúp robot nhận diện và tương tác với môi trường xung quanh.

Trong những năm gần đây, hai kiến trúc phát hiện đối tượng có ảnh hưởng sâu rộng và hình thành nền tảng cho nhiều hệ thống thị giác máy tính hiện đại là Mạng Nơ-ron Tích chập (Convolutional Neural Networks - CNN) và You Only Look Once (YOLO).

Mạng Nơ-ron Tích chập (CNN) sử dụng phép tích chập như một cơ chế cốt lõi để trích xuất và học các đặc trưng không gian của hình ảnh. Bằng cách áp dụng các bộ

lọc trượt trên toàn bộ ảnh, CNN có khả năng phát hiện các đặc trưng từ thấp đến cao, từ cạnh, góc, cho đến các cấu trúc phức tạp. Từ kiến trúc nền tảng này, nhiều biến thể đã được phát triển nhằm cải thiện tốc độ và độ chính xác, tiêu biểu như R-CNN, Fast R-CNN hay Mask R-CNN, góp phần nâng cao hiệu quả trong các nhiệm vụ định vị và phân loại đối tượng.

YOLO (You Only Look Once) là một trong những mô hình phát hiện đối tượng thời gian thực nổi bật nhất. Được đề xuất lần đầu vào năm 2015, YOLO mang tính đột phá khi tiếp cận bài toán phát hiện đối tượng theo cách hoàn toàn mới: thay vì xử lý theo từng vùng như các mô hình truyền thống, YOLO dự đoán trực tiếp bounding boxes và nhãn lớp chỉ qua một lần quan sát toàn bộ ảnh. Nhờ đó, YOLO đạt được tốc độ xử lý rất cao trong khi vẫn đảm bảo độ chính xác đáng kể, và nhanh chóng trở thành lựa chọn phổ biến trong nhiều ứng dụng thực tế.

Tại Việt Nam, các nghiên cứu đã khai thác mô hình YOLO trong các bài toán như nhận dạng biển số xe và giám sát giao thông, cho thấy hiệu quả trong các hệ thống quản lý và kiểm soát thông minh. Đồng thời, nhiều công trình cũng tiến hành so sánh giữa YOLO và các thuật toán khác như SSD nhằm xác định giải pháp tối ưu cho từng nhu cầu ứng dụng cụ thể.

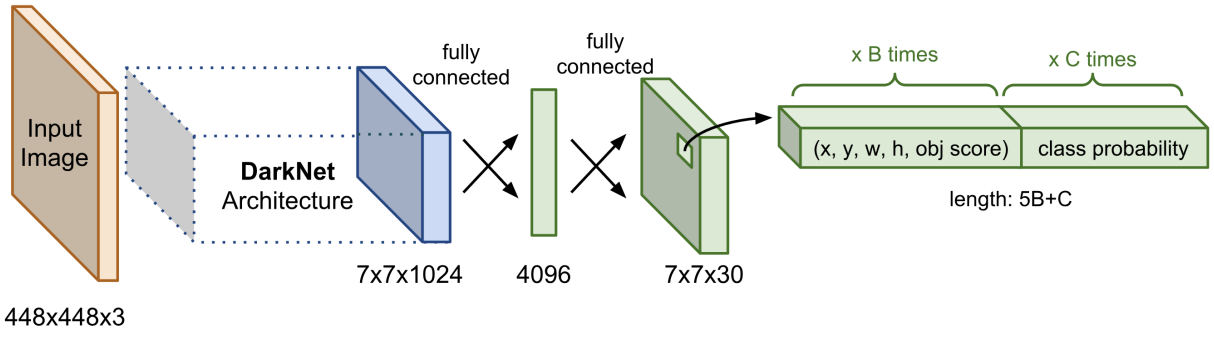
2.6 GIỚI THIỆU MÔ HÌNH YOLO

YOLO (You Only Look Once) là một mô hình phát hiện đối tượng dựa trên mạng nơ-ron tích chập (CNN), được thiết kế để thực hiện nhận dạng, phân loại và định vị đối tượng trong ảnh một cách nhanh chóng và hiệu quả. Không giống như các phương pháp truyền thống chia bài toán thành nhiều giai đoạn (như đề xuất vùng và phân loại), YOLO xử lý toàn bộ ảnh chỉ trong một lần duy nhất, từ đó mang lại tốc độ vượt trội. [9] Bản chất "nhìn một lần" của YOLO giúp mô hình đạt tốc độ xử lý thời gian thực mà vẫn duy trì độ chính xác cao, khiến nó trở thành một trong những kiến trúc phổ biến nhất trong lĩnh vực phát hiện đối tượng cho các ứng dụng như giám sát thông minh, xe tự hành và phân tích video. Về mặt kiến trúc, YOLO được cấu trúc thành ba thành phần chính: Backbone, Neck, và Head, mỗi phần đảm nhiệm một vai trò quan trọng trong quá trình

phát hiện đối tượng:

- **Backbone (Mạng trích xuất đặc trưng):** YOLO sử dụng một mạng nơ-ron tích chập sâu (CNN) làm nền tảng để trích xuất đặc trưng từ ảnh đầu vào. Backbone học các thông tin quan trọng như cạnh, hình dạng, họa tiết và cấu trúc đối tượng, tạo nên nền tảng cho quá trình phát hiện chính xác.
- **Neck (Phần kết nối đặc trưng):** Bộ phận này thường tích hợp các kiến trúc như FPN (Feature Pyramid Network) hoặc PANet (Path Aggregation Network). Neck có nhiệm vụ kết hợp và khuếch tán thông tin từ nhiều tầng của Backbone, giúp mô hình duy trì khả năng phát hiện tốt đối với các đối tượng ở nhiều kích thước khác nhau, từ rất nhỏ đến rất lớn.
- **Head (Phần dự đoán đầu ra):** Đây là nơi thực hiện các phép dự đoán cuối cùng, bao gồm tọa độ hộp giới hạn (bounding boxes), điểm tin cậy (confidence scores) và xác suất thuộc lớp đối tượng. Head tổng hợp thông tin từ Backbone và Neck để đưa ra kết quả phát hiện hoàn chỉnh.

Nhờ sự phối hợp hiệu quả giữa ba thành phần này, YOLO đạt được tốc độ xử lý nhanh trong khi vẫn giữ được độ chính xác cao, trở thành một trong những kiến trúc phát hiện đối tượng hiệu quả nhất hiện nay.



Hình 2.6.6: Sơ đồ kiến trúc mạng YOLO

2.6.1 Nguyên lý hoạt động của YOLO

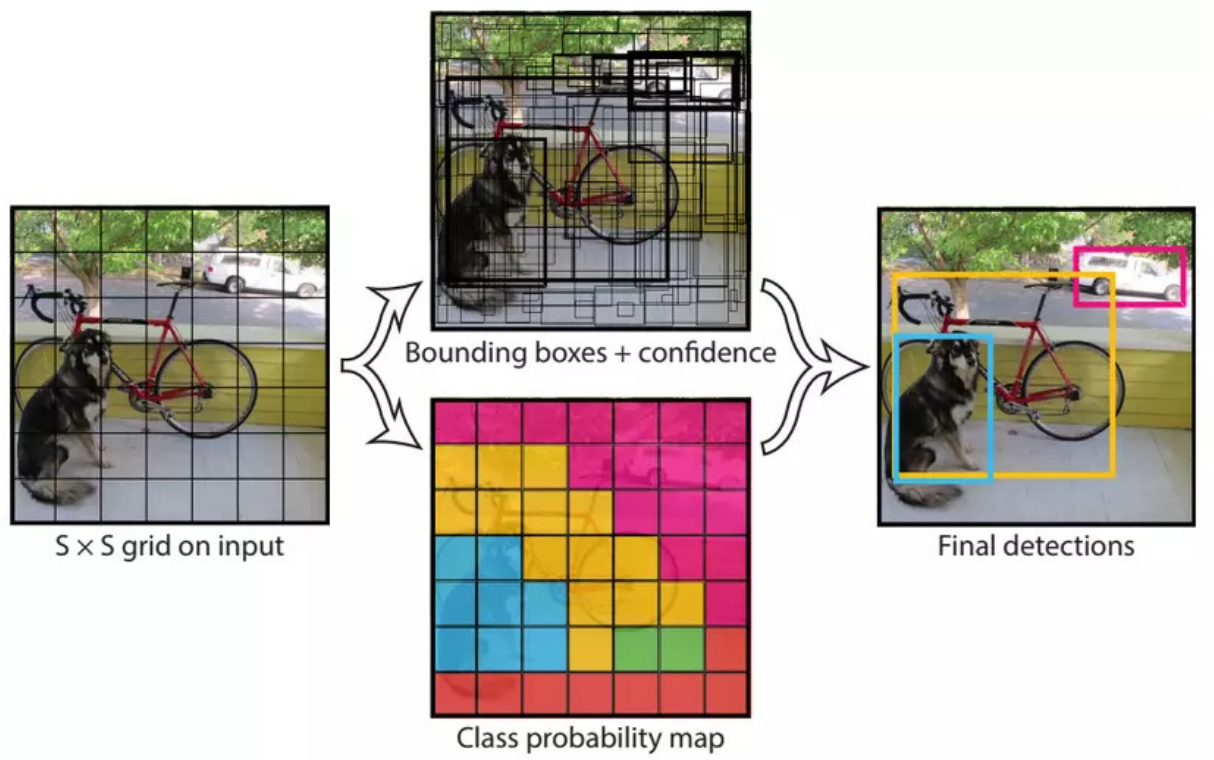
YOLO hoạt động theo 4 bước chính:

- **Chia ảnh thành lưới (grid) S x S:** Ảnh đầu vào được chia thành một lưới gồm S hàng × S cột ô (grid-cells). Mỗi ô "chịu trách nhiệm" phát hiện các đối tượng mà "tâm" của hộp giới hạn (bounding box) rơi vào ô đó. Ý tưởng là phân vùng ảnh để mỗi phần nhỏ có thể dự đoán xem có đối tượng hay không, thay vì lặp toàn bộ ảnh nhiều lần.

- **Trích xuất đặc trưng ảnh với mạng CNN (backbone + feature layers):** YOLO sử dụng một mạng nơ-ron tích chập (CNN) để "đọc" toàn bộ ảnh – các lớp convolution + pooling + ... trích xuất đặc trưng (features) từ ảnh. Các lớp cuối (fully connected / detection layers) sau khi feature extraction sẽ dùng để dự đoán bounding boxes + nhãn + độ tin cậy.

- **Dự đoán nhiều bounding boxes + xác suất + lớp đối tượng ở mỗi ô lưới:** Mỗi ô (grid cell) dự đoán B hộp giới hạn (bounding boxes). Mỗi hộp gồm các thông số: tâm x, y; chiều rộng w, chiều cao h (thường được chuẩn hóa), + một "điểm tin cậy (confidence score)" biểu thị: khả năng có đối tượng + độ tin cậy vị trí. Ngoài ra, mỗi ô cũng dự đoán xác suất (conditional class probabilities) cho mỗi lớp đối tượng mà ảnh đó có thể chứa.

- **Kết hợp kết quả và loại bỏ dư thừa (Non-Maximum Suppression – NMS):** Vì nhiều hộp có thể "đụng chồng" nhau (ví dụ cùng dự đoán một đối tượng), sau khi mạng đưa ra tất cả dự đoán, YOLO sử dụng kỹ thuật NMS để giữ lại hộp có "điểm tin cậy tốt nhất" và loại bỏ các hộp dư chồng lặp. Kết quả cuối: một danh sách các hộp (bounding boxes), mỗi hộp có nhãn lớp đối tượng và độ tin cậy — tương ứng với những đối tượng được phát hiện trong ảnh.



Hình 2.6.7: Cơ chế hoạt động của YOLO

2.6.2 Hàm mất mát (Loss Function) trong YOLO

Trong YOLO, hàm mất mát (loss function) được xây dựng từ sự khác biệt giữa dự đoán của mô hình và nhãn thực tế. Tổng độ lỗi là sự kết hợp của ba thành phần chính, mỗi thành phần phản ánh một khía cạnh quan trọng trong quá trình phát hiện đối tượng:

1. **Classification Loss - Sai số phân loại:** Đo mức độ chính xác khi mô hình dự đoán lớp của đối tượng trong mỗi ô lưới. Mục tiêu là đảm bảo mô hình không chỉ phát hiện được vật thể, mà còn nhận diện đúng loại của nó.

2. **Localization Loss – Sai số định vị bounding box:** Đánh giá độ lệch giữa bounding box dự đoán và bounding box thật dựa trên bốn tham số: vị trí tâm (x, y) và kích thước (w, h). Thành phần này giúp mô hình học cách khoanh vùng đối tượng chính xác hơn.

3. **Confidence Loss – Sai số về mức độ tin cậy:** Phản ánh sự khác biệt giữa "mức độ chắc chắn" mà mô hình cho rằng ô lưới chứa một vật thể và giá trị nhãn thực tế. Đây là yếu tố quan trọng giúp YOLO biết được ô nào nên dự đoán và ô

nào nên bỏ qua.

Classification Loss:

Classification loss là sai số phản ánh mức độ chính xác khi mô hình dự đoán lớp của đối tượng. Thành phần này chỉ được tính cho những ô lưới thật sự chứa object, còn các ô không có vật thể sẽ được bỏ qua để tránh làm nhiễu quá trình học. Classification loss được xác định theo công thức sau:

$$L_{classification} = \sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \tag{2.6.4}$$

Trong đó:

- I_i^{obj}: Biến chỉ báo, bằng 1 nếu ô lưới i chứa object, ngược lại bằng 0.
- $\hat{p}_i(c)$: Xác xuất có điều kiện của lớp c tại ô vuông tương ứng mà mô hình dự đoán.

Localization Loss:

Localization loss là thành phần đo sai số vị trí và kích thước của bounding box mà mô hình dự đoán. Nó so sánh tọa độ tâm cùng chiều rộng và chiều cao của bounding box dự đoán với giá trị thực tế (ground truth) trong dữ liệu huấn luyện. Một điểm quan trọng là các giá trị này không được tính trực tiếp theo kích thước ảnh gốc, mà phải được chuẩn hoá về khoảng [0, 1] dựa trên kích thước của ô lưới và vị trí tương đối của bounding box. Việc chuẩn hóa giúp mô hình học ổn định hơn, hội tụ nhanh hơn và đạt độ chính xác cao hơn. Localization loss được tính bằng tổng sai số giữa độ tâm (x,y) và kích thước (w,h) của bounding box dự đoán so với ground truth. Ở mỗi ô lưới chứa đối tượng, YOLO chọn một bounding box có IOU cao nhất với ground truth để chịu trách nhiệm dự đoán. Sai số được tính dựa trên bounding box được chọn này. Công thức tính Localization loss như sau:

$$L_{localization} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \tag{2.6.5}$$

Trong đó:

- $_{coord}$: Hệ số điều chỉnh, thường được đặt giá trị cao hơn để nhấn mạnh tầm quan trọng của localization loss.
- 1_{ij}^{obj}: Biến chỉ báo, bằng 1 nếu ô lưới i chứa object và bounding box j chịu trách nhiệm dự đoán, ngược lại bằng 0.
- x_i, y_i, w_i, h_i: Tọa độ tâm và kích thước thực tế của bounding box.
- $\hat{x_i}, \hat{y_i}, \hat{w_i}, \hat{h_i}$: Tọa độ tâm và kích thước dự đoán của bounding box.

Confidence Loss:

Confidence loss đo mức độ chênh lệch giữa "độ tin cậy" mà mô hình gán cho một bounding box (khả năng ô lưới đó chứa object) và giá trị nhãn thực tế. Khác với hai thành phần loss còn lại, confidence loss được tính cho tất cả các ô lưới, bao gồm cả ô có đối tượng và ô không có đối tượng. Nhờ đó, mô hình học được cách phân biệt đâu là vùng chứa object thật và đâu là vùng nền (background), giúp giảm dự đoán sai và tăng độ chính xác tổng thể.

$$L_{confidence} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[1_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{noobj} 1_{ij}^{noobj} (C_i - \hat{C}_i)^2 \right] \tag{2.6.6}$$

Trong đó:

- λ_{noobj}: Hệ số điều chỉnh quan trọng được sử dụng để kiểm soát trọng số của phần lỗi liên quan đến các bounding box không chứa đối tượ (no-object).
- 1_{ij}^{noobj}: Cho biết bounding box thứ của cell i không chứa đối tượng.
- C_i: Độ tin cậy thực tế (1 nếu có đối tượng, 0 nếu không có).
- $\hat{C}_i$: Độ tin cậy dự đoán của bounding box.

Tổng hợp hàm mất mát trong YOLO:

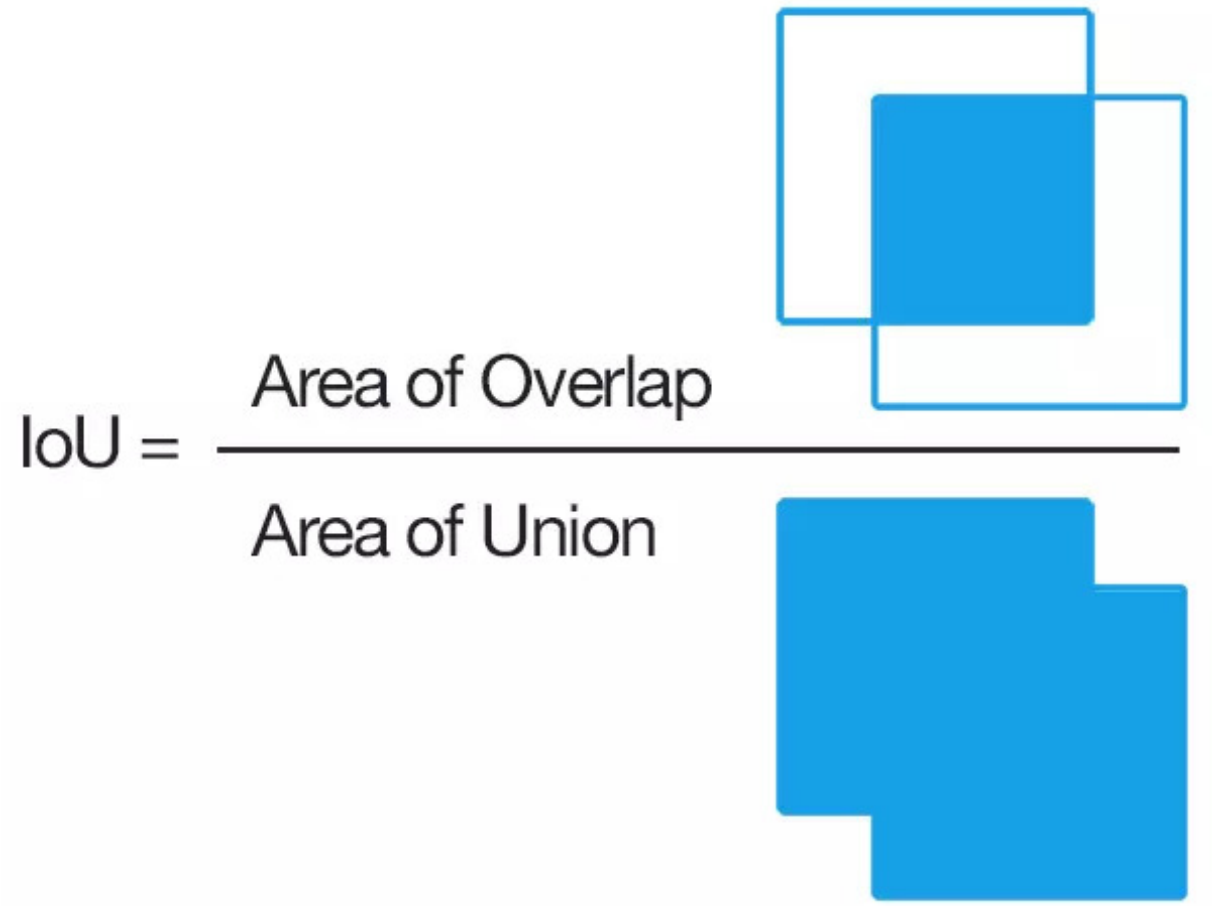
$$L_{total} = L_{classification} + L_{localization} + L_{confidence} \tag{2.6.7}$$

2.6.3 Các chỉ số đánh giá hiệu suất của YOLO

Trong bài toán phát hiện đối tượng, ba chỉ số quan trọng thường được sử dụng để đánh giá hiệu suất mô hình gồm Mean Average Precision (mAP), Average Precision (AP) và Intersection over Union (IoU). **Mean Average Precision (mAP):** mAP đánh giá hiệu suất tổng thể của mô hình bằng cách lấy trung bình giá trị AP trên toàn bộ các lớp đối tượng. Để tính mAP, trước hết ta xác định AP của từng lớp dựa trên tích phân của đường cong Precision–Recall. Sau đó, mAP được tính bằng trung bình cộng của các AP này. Giá trị mAP càng cao chứng tỏ mô hình đạt được độ chính xác (precision) và độ thu hồi (recall) tốt trên toàn bộ tập đối tượng. **Average Precision (AP):** AP phản ánh hiệu suất của mô hình tại các mức độ thu hồi khác nhau thông qua quan hệ giữa hai đại lượng:

- Precision (Độ chính xác): TP/(TP + FP)
- Recall (Độ thu hồi): TP/(TP + FN)

AP được tính bằng diện tích dưới đường cong Precision-Recall (PR curve), thể hiện sự đánh đổi giữa khả năng phát hiện đúng (precision) và khả năng tìm được nhiều đối tượng nhất (recall). **Intersection over Union (IoU):** IoU đo mức độ trùng khớp giữa bounding box dự đoán và bounding box ground truth. Chỉ số này được tính bằng tỉ lệ giữa diện tích vùng giao nhau và diện tích vùng hợp nhất của hai bounding box. Giá trị IoU càng cao cho thấy mô hình định vị đối tượng càng chính xác. Công thức tính như hình 2.6.8 dưới đây:



Hình 2.6.8: Cách tính chỉ số IOU

2.6.4 NON-MAXIMUM SUPPRESSION

Trong giai đoạn suy luận, YOLO thường tạo ra nhiều bounding box trùng lặp, đặc biệt tại những khu vực có mật độ đối tượng cao hoặc khi các ô lưới lân cận cùng dự đoán một vật thể. Việc xuất hiện quá nhiều bounding box chồng chéo không chỉ gây dư thừa mà còn làm giảm chất lượng dự đoán. Để khắc phục vấn đề này, YOLO sử dụng kỹ thuật Non-Maximum Suppression (NMS). Phương pháp này sàng lọc và loại bỏ các bounding box kém quan trọng, chỉ giữ lại hộp có độ tin cậy cao nhất trong số các hộp chồng lặp lên nhau. Nhờ đó, mô hình tránh được việc "đếm" một đối tượng nhiều lần và tập trung vào các dự đoán chính xác nhất, giúp kết quả nhận dạng trở nên rõ ràng và đáng tin cậy hơn.

Hình 2.6.9: Kết quả sau khi áp dụng Non-Maximum Suppression

Các bước của Non-Maximum Suppression bao gồm:

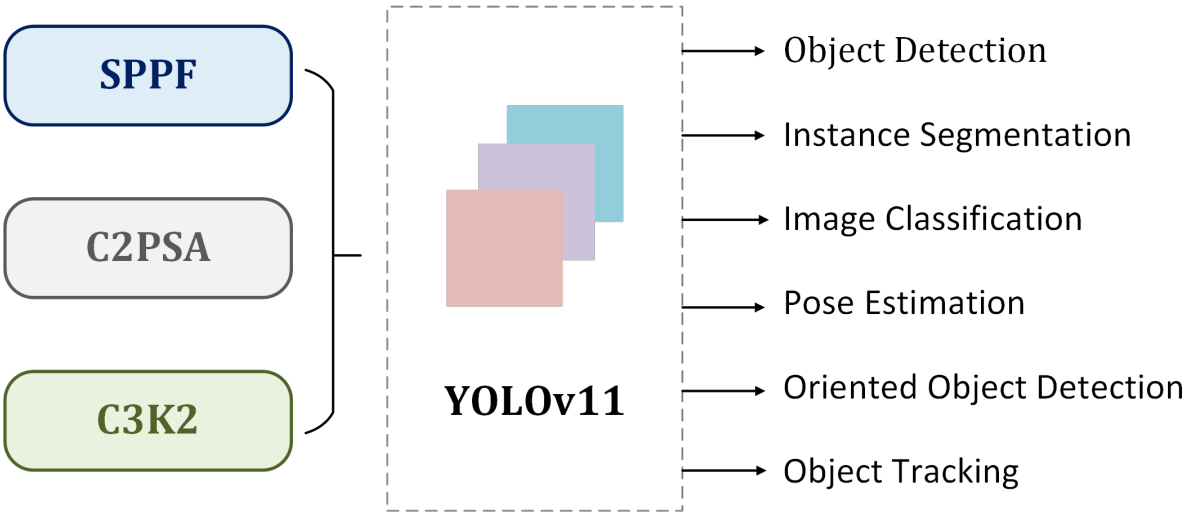
- Bước 1: Loại bỏ các bounding box có xác suất chứa vật thể nhỏ hơn một ngưỡng đã cho, thường là 0.5.
- Bước 2: Chọn bounding box có độ tin cậy cao nhất.
- Bước 3: Tính toán IoU giữa bounding box đã chọn và các bounding box còn lại. Loại bỏ những hộp có IoU lớn hơn một ngưỡng nhất định (ví dụ 0.4) để tránh chồng lấp.

2.7 TỔNG QUAN VỀ MÔ HÌNH YOLOV11

YOLO11 là thế hệ mới nhất trong chuỗi mô hình YOLO do Ultralytics phát triển, hướng tới bài toán phát hiện đối tượng theo thời gian thực với hiệu năng vượt trội. Kế thừa và mở rộng những thành tựu của các phiên bản tiền nhiệm, YOLO11 được trang bị nhiều cải tiến quan trọng về kiến trúc mạng cũng như chiến lược huấn luyện, qua đó nâng cao đồng thời độ chính xác, tốc độ suy luận và hiệu quả tính toán. Nhờ tính linh hoạt và khả năng thích ứng cao, YOLO11 phù hợp với nhiều bài toán thị giác máy tính trong các kịch bản ứng dụng thực tế.

2.7.1 Kiến trúc mô hình YOLOv11

Kế thừa nền tảng kiến trúc đã được khẳng định, YOLOv11 tiếp tục phát triển và hoàn thiện những thành quả của YOLOv8 thông qua các cải tiến về cấu trúc mạng và chiến lược tối ưu tham số, từ đó nâng cao đáng kể hiệu quả và độ chính xác trong bài toán phát hiện đối tượng.



Hình 2.7.10: Sơ đồ kiến trúc mạng YOLOv11

Các khối trong kiến trúc YOLOv11 bao gồm:

- **SPFF (Spatial Pyramid Pooling - Fast):** Khối SPPF được sử dụng để mở rộng vùng cảm thụ (receptive field) của mạng bằng cách áp dụng các phép pooling với kích thước khác nhau trên cùng một đặc trưng đầu vào. Nhờ đó, mô hình có thể khai thác thông tin ngữ cảnh đa tỷ lệ mà không làm tăng đáng kể chi phí tính toán, giúp cải thiện khả năng phát hiện các đối tượng có kích thước khác nhau.
- **C2PSA (Cross Stage Partial with Self-Attention):** C2PSA là khối đặc trưng kết hợp giữa cơ chế Cross Stage Partial (CSP) và Self-Attention. Khối này giúp tăng cường khả năng học mối quan hệ không gian – ngữ nghĩa giữa các vùng trong ảnh, đồng thời giảm số lượng tham số và chi phí tính toán. Nhờ đó, mô hình có thể tập trung tốt hơn vào các vùng quan trọng của đối tượng.
- **C3K2:** C3K2 là một biến thể của khối C3, sử dụng các lớp tích chập với kernel kích thước nhỏ (ví dụ 3x3) và cấu trúc residual để trích xuất đặc trưng hiệu quả.

Khối này giúp cân bằng giữa độ sâu mạng, khả năng biểu diễn đặc trưng và tốc độ suy luận, đặc biệt phù hợp cho các tác vụ phát hiện đối tượng thời gian thực.

- **Các khối đặc trưng nhiều màu ở trung tâm hình:** Các khối này biểu diễn các feature maps ở nhiều mức không gian khác nhau, tương ứng với các tầng đặc trưng đa tỷ lệ được trích xuất từ backbone/neck. Việc kết hợp các feature maps này cho phép mô hình phát hiện hiệu quả cả đối tượng nhỏ, trung bình và lớn.

2.7.2 Tính năng nổi bật của mô hình YOLOv11

YOLOv11 được phát triển như một bước tiến quan trọng trong dòng mô hình YOLO, tập trung đồng thời vào độ chính xác, tốc độ xử lý và hiệu quả tính toán. Những đặc điểm nổi bật của mô hình có thể được tóm lược như sau:

- **Trích xuất đặt trưng nâng cao:** YOLOv11 áp dụng kiến trúc backbone và neck được cải tiến, cho phép khai thác đặc trưng hình ảnh ở nhiều mức không gian và ngữ nghĩa khác nhau. Nhờ đó, mô hình nâng cao khả năng biểu diễn đặc trưng, đặc biệt hiệu quả trong việc phát hiện các đối tượng nhỏ, chồng lấp hoặc xuất hiện trong bối cảnh phức tạp.

- **Tối ưu hóa hiệu suất và tốc độ suy luận:** Với thiết kế kiến trúc tinh gọn cùng quy trình huấn luyện được tối ưu hóa, YOLOv11 đạt tốc độ xử lý vượt trội trong khi vẫn duy trì sự cân bằng hợp lý giữa độ chính xác và hiệu năng. Điều này giúp mô hình đáp ứng tốt các yêu cầu của các ứng dụng thời gian thực.

- **Độ chính xác cao với số lượng tham số giảm:** Nhờ các cải tiến trong thiết kế mô hình, YOLOv11m đạt giá trị mAP cao hơn trên bộ dữ liệu COCO, đồng thời giảm khoảng 22% số lượng tham số so với YOLOv8m. Sự tối ưu này giúp nâng cao hiệu quả tính toán, giảm yêu cầu tài nguyên mà không làm suy giảm chất lượng dự đoán.

- **Khả năng thích ứng linh hoạt trong nhiều môi trường triển khai:** YOLOv11 có thể được triển khai hiệu quả trên nhiều nền tảng khác nhau, từ các thiết bị biên (edge devices), hệ thống nhúng, đến các môi trường đám mây và hạ tầng tăng tốc

GPU của NVIDIA. Tính linh hoạt này giúp mô hình dễ dàng tích hợp vào các hệ thống ứng dụng thực tế.

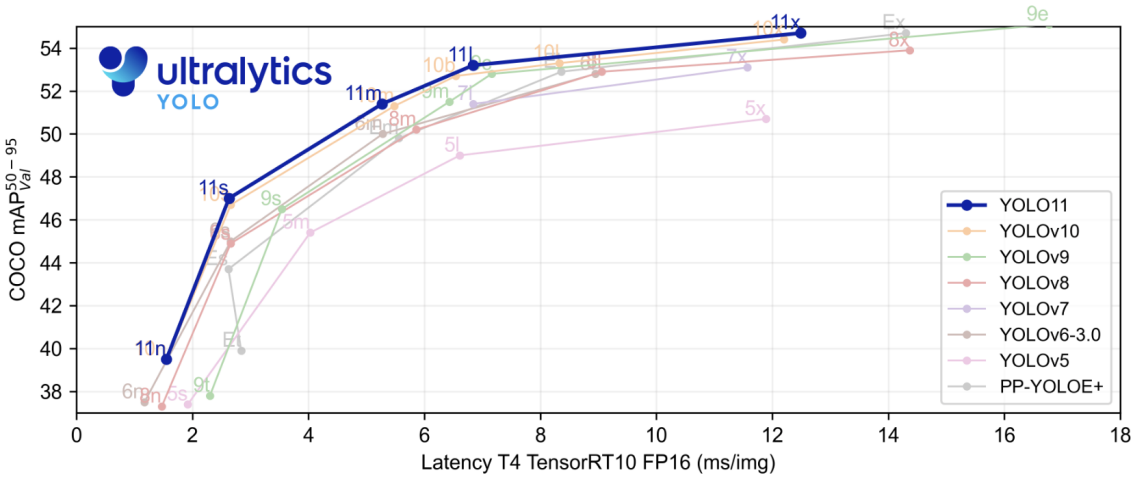
- **Hỗ trợ đa dạng các tác vụ thị giác máy tính:** Không chỉ giới hạn ở bài toán phát hiện đối tượng, YOLOv11 còn được thiết kế để hỗ trợ nhiều nhiệm vụ thị giác máy tính khác nhau như phân đoạn ảnh, phân loại hình ảnh, ước tính tư thế, phát hiện đối tượng theo hướng (oriented object detection). Nhờ đó, phạm vi ứng dụng của YOLOv11 được mở rộng đáng kể trong các bài toán thực tiễn.

Hiệu suất của mô hình YOLOv11 YOLOv11 - phiên bản mới nhất trong họ mô hình YOLO - đã được Ultralytics tiến hành đánh giá và so sánh hiệu suất với các phiên bản tiền nhiệm từ YOLOv5 đến YOLOv10. Kết quả thực nghiệm cho thấy YOLOv11 đạt được những cải tiến rõ rệt cả về độ chính xác lẫn tốc độ suy luận, qua đó khẳng định ưu thế vượt trội của kiến trúc mới.

Cụ thể, YOLOv11x đạt giá trị khoảng 54,5% $mAP50_{50-95}$ với độ trễ suy luận chỉ 13 ms, vượt qua toàn bộ các phiên bản YOLO trước đó trong cả hai tiêu chí độ chính xác và hiệu năng. Điều này cho thấy mô hình có khả năng xử lý các bài toán phát hiện đối tượng phức tạp trong thời gian thực với độ tin cậy cao.

Đối với YOLOv11m, mô hình mang lại mức độ chính xác tương đương hoặc tiệm cận với các biến thể kích thước lớn của các thế hệ YOLO trước, nhưng yêu cầu tài nguyên tính toán thấp hơn đáng kể. Đặc điểm này giúp YOLOv11m trở thành lựa chọn cân bằng giữa hiệu năng và chi phí triển khai.

Trong khi đó, YOLOv11s hướng đến các hệ thống yêu cầu độ trễ cực thấp. Mô hình đạt khoảng 47% mAP_{50-95} trong khoảng độ trễ từ 2-6 ms, cho phép triển khai hiệu quả trong các ứng dụng thời gian thực trên thiết bị biên mà vẫn duy trì mức độ chính xác chấp nhận được. Nhìn chung, các kết quả đánh giá cho thấy YOLOv11 không chỉ cải thiện hiệu suất so với các phiên bản trước, mà còn mở rộng khả năng ứng dụng trong nhiều kịch bản khác nhau, từ hệ thống nhúng hạn chế tài nguyên đến các nền tảng tính toán hiệu năng cao.



Hình 2.7.11: Hiệu suất của mô hình yolov11 so với các phiên bản trước

Ngoài ra, đường cong cải tiến của YOLOv11 cho thấy khả năng mở rộng (scaling) vượt trội giữa các biến thể của mô hình, cho phép khai thác tài nguyên tính toán một cách hiệu quả và linh hoạt hơn so với các thế hệ trước. Đặc tính này giúp các phiên bản YOLOv11 duy trì sự cân bằng tối ưu giữa độ chính xác và hiệu năng khi thay đổi quy mô mô hình. Những kết quả đạt được đã khẳng định YOLOv11 là một bước tiến đáng kể trong lĩnh vực phát hiện đối tượng theo thời gian thực, đáp ứng hiệu quả các yêu cầu ngày càng khắt khe của các ứng dụng thị giác máy tính hiện đại.

2.8 Công cụ mô phỏng giao thông: SUMO (Simulation of Urban Mobility)

2.8.1 Giới thiệu SUMO

2.8.2 Ưu điểm và hạn chế của SUMO trong nghiên cứu mô phỏng và thực tế giao thông đô thị

2.9 TỔNG QUAN VỀ MÔ HÌNH LSTM

Chương 3

THIẾT KẾ HỆ THỐNG

3.1 YÊU CẦU CỦA HỆ THỐNG

Hệ thống được đề xuất hướng tới việc mô phỏng, dự báo và điều tiết giao thông dựa trên dữ liệu hình ảnh snapshot thu thập từ camera giao thông, kết hợp các kỹ thuật học sâu và mô phỏng giao thông vi mô. Để đáp ứng mục tiêu nghiên cứu và đảm bảo khả năng triển khai thực tế, hệ thống cần thỏa mãn các yêu cầu chức năng và phi chức năng sau.

Về giai đoạn thu thập dữ liệu, hệ thống cần đảm bảo các yêu cầu sau:

- Tiếp nhận được dữ liệu hình ảnh giao thông dạng snapshot từ các camera giao thông đặt tại các nút giao.
- Các ảnh này phải được quản lý, lưu trữ và gắn nhãn thời gian rõ ràng.

Về giai đoạn tiền xử lý dữ liệu, nhận diện và đếm phương tiện, hệ thống cần đáp ứng các yêu cầu sau:

- Các ảnh đầu vào cần được thực hiện các bước xử lý ảnh cơ bản như tăng độ phân giải, giảm nhiễu và chuẩn hóa kích thước.
- Hệ thống tích hợp mô hình YOLOv11 nhằm nhận diện các phương tiện giao thông

chính trên từng ảnh snapshot, trong đó các phương tiện được quy ước và phân loại thành hai nhóm: xe hai bánh (xe máy) và xe bốn bánh (xe hơi).

- Kết quả phát hiện bao gồm số lượng phương tiện theo từng loại tại mỗi thời điểm và tại từng địa điểm giám sát cụ thể, trong đó "địa điểm" được hiểu là vị trí camera đại diện cho một khu vực giao thông trên bản đồ.
- Lưu trữ kết quả đếm phương tiện với timestamp và vị trí tương ứng để phục vụ cho các bước xử lý tiếp theo.
- Đảm bảo độ chính xác cao trong việc nhận diện và đếm phương tiện, với sai số không vượt quá 5% so với thực tế.

Tiếp đến là yêu cầu về mô phỏng giao thông:

- None

Đối với giai đoạn xây dựng chuỗi thời gian và dự báo lưu lượng giao thông, hệ thống cần thỏa mãn các yêu cầu sau:

- None

Đối với giai đoạn tích hợp và điều khiển luồng giao thông, hệ thống cần đáp ứng các yêu cầu sau:

- Hệ thống tích hợp các giá trị dự báo từ LSTM vào môi trường mô phỏng giao thông SUMO, trong đó lưu lượng phương tiện, tốc độ dòng xe hoặc phân bố phương tiện tại các nút giao được điều chỉnh tương ứng với trạng thái giao thông dự kiến.
- None

Cuối cùng, về giai đoạn trực quan hóa và phân tích, hệ thống cần:

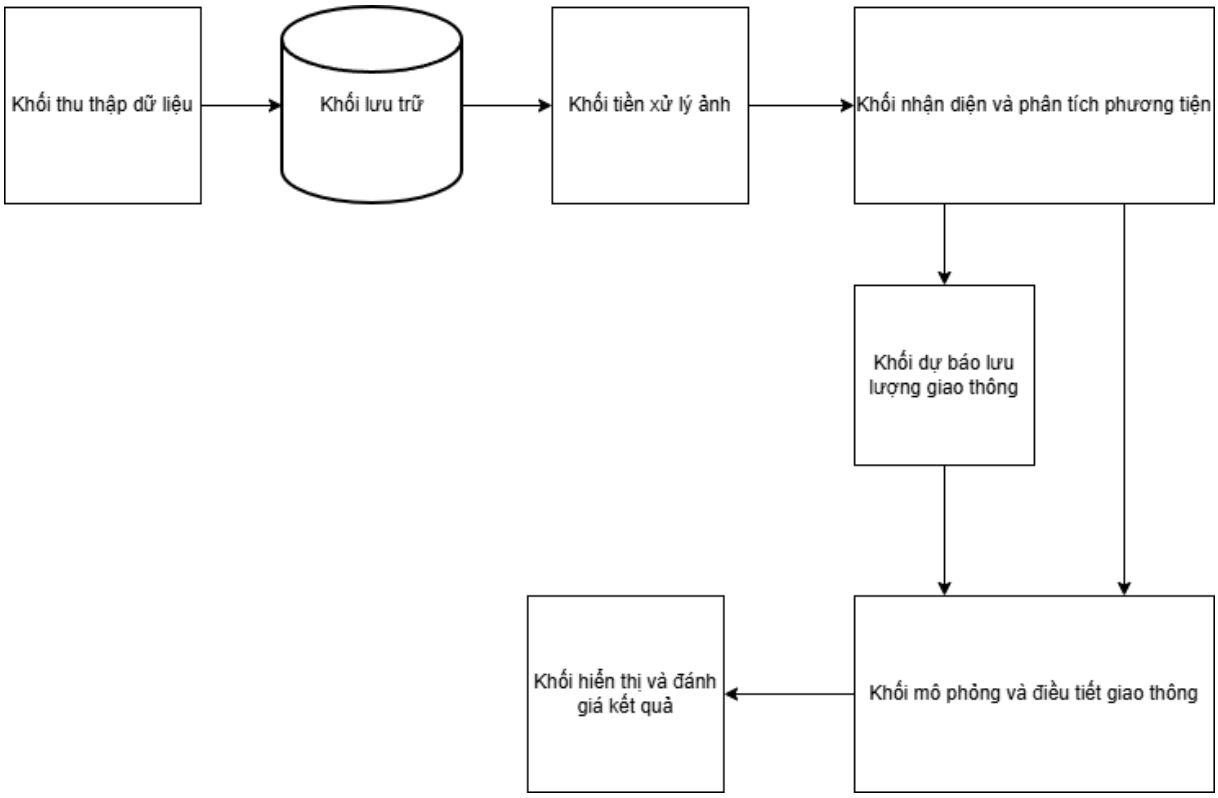
- Cung cấp giao diện trực quan để hiển thị kết quả nhận diện, đếm phương tiện, dự báo lưu lượng và mô phỏng giao thông.

- Hỗ trợ các biểu đồ, bản đồ nhiệt và các công cụ phân tích để người dùng có thể dễ dàng hiểu và đánh giá tình hình giao thông.

- So sánh hiệu quả giữa việc xử dụng chiến lược điều tiết và không sử dụng để đánh giá hiệu quả của các biện pháp điều tiết dựa trên các chỉ số: thời gian di chuyển trung bình, thời gian đợi, chiều dài hàng đợi.

3.2 KIẾN TRÚC HỆ THỐNG

3.2.1 Sơ đồ khối hệ thống



Hình 3.2.1: Sơ đồ khối kiến trúc hệ thống

Chức năng từng khối:

- **Khối thu thập dữ liệu:** Khối thu thập dữ liệu chịu trách nhiệm tiếp nhận các hình ảnh snapshot từ nguồn mở Open Street Map (OSM). Trong đề tài này, dữ liệu không được thu thập trực tiếp từ hệ thống camera vật lý mà được lấy từ nguồn dữ

liệu mở, nơi các camera giao thông đã được thiết lập và công bố sẵn cho mục đích quan sát và tham khảo. Quá trình thu thập dữ liệu được thực hiện thông qua truy suất tự động (web scraping) để tải về các hình ảnh snapshot tại các thời điểm xác định (khoảng 12 giây một ảnh).

- **Khối lưu trữ:** Khối lưu trữ đóng vai trò quản lý và tổ chức toàn bộ dữ liệu hình ảnh thu thập được cũng như các dữ liệu trung gian phát sinh trong quá trình xử lý. Các ảnh snapshot thu thập từ camera phía OSM được lưu trữ dưới dạng cấu trúc thư mục phân cấp trên hệ thống lưu trữ tệp tin đám mây Google Drive. Mỗi ảnh được gắn nhãn thời gian rõ ràng và sắp xếp theo từng thư mục tương ứng với địa điểm giám sát, nhằm đảm bảo tính nhất quán và thuận tiện cho việc truy xuất, xử lý và phân tích về sau. Bên cạnh đó, các ảnh kết quả sau khi nhận diện và đếm phương tiện, dữ liệu thống kê mật độ giao thông theo thời gian, cũng như dữ liệu đầu vào và đầu ra của mô hình LSTM đều được lưu trữ tập trung tại khối này.

- **Khối tiền xử lý ảnh:** Khối tiền xử lý ảnh đảm nhiệm việc thực hiện các phép biến đổi cần thiết nhằm cải thiện chất lượng dữ liệu hình ảnh trước khi đưa vào mô hình nhận diện. Nguyên nhân là do các ảnh snapshot thu thập từ camera giao thông công khai trên mạng thường có chất lượng không đồng đều, chịu ảnh hưởng bởi nhiều yếu tố như độ phân giải thấp, nhiễu ảnh, điều kiện ánh sáng không ổn định và góc chụp chưa tối ưu. Do đó, khối này áp dụng một số kỹ thuật xử lý ảnh cơ bản, bao gồm tăng cường độ phân giải (super-resolution), lọc nhiễu (denoising), chuẩn hóa kích thước ảnh (resizing) và chia nhỏ ảnh nhằm đảm bảo dữ liệu đầu vào cho mô hình nhận diện đạt chất lượng tốt hơn và có tính nhất quán, dễ dàng nhận diện.

- **Khối nhận diện và phân tích phương tiện:** Khối nhận diện và phân tích phương tiện sử dụng mô hình học sâu YOLOv11 để thực hiện phát hiện và phân loại các phương tiện giao thông xuất hiện trong từng ảnh snapshot. Mô hình được khởi tạo từ trọng số huấn luyện sẵn (pre-trained) trên các tập dữ liệu lớn và đa dạng, nhờ đó có khả năng nhận diện hiệu quả các loại phương tiện phổ biến. Tuy nhiên, nhằm đơn giản hóa bài toán và phù hợp với mục tiêu nghiên cứu, các phương tiện được quy ước và gom nhóm thành hai lớp chính, bao gồm xe hai bánh (đại diện cho xe máy) và xe bốn bánh (đại diện cho xe hơi). Bên cạnh đó, do chất lượng hình

ảnh thu thập từ camera giao thông còn hạn chế và các bước tiền xử lý không thể khắc phục hoàn toàn các yếu tố bất lợi như góc quay cao, góc quay xiên hoặc mật độ giao thông lớn, một số phương tiện—đặc biệt là xe máy—có thể bị che khuất hoặc chồng chéo, dẫn đến độ chính xác nhận diện suy giảm. Để khắc phục vấn đề này, khối xử lý còn tích hợp các thuật toán xử lý ảnh bổ sung như Polygon Fill, Rectangle Fill, Pixel-wise Logic và Non-zero Pixel Count nhằm tính toán tỷ lệ che phủ của xe máy trong các vùng quan tâm. Trên cơ sở đó, số lượng xe máy trong từng vùng được ước lượng, góp phần nâng cao độ chính xác tổng thể của quá trình đếm phương tiện.

- **Khối dự báo lưu lượng giao thông:** None
- **Khối mô phỏng và điều tiết giao thông:** None
- **Khối hiển thị và đánh giá kết quả:** None

3.3 Upscale ảnh bằng Real ESRGAN

TÀI LIỆU THAM KHẢO

[1] OECD, "Traffic management and congestion control," *OECD Transport Outlook*, 2019.

[2] Y. Zhang, S. Wang, and G. Chen, "Deep learning for traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 4, pp. 1044–1055, 2018.

[3] P. T. N. T. Hải, *Giáo trình Xử Lý Ảnh Số*. Nhà xuất bản Thông tin và Truyền thông hoặc NXB Bách khoa Hà Nội, 2015.

[4] GeeksforGeeks. What is image processing? Accessed: 2025-10-21. [Online]. Available: https://www.geeksforgeeks.org/electronics-engineering/what-is-image-processing/

[5] v. A. C. Ian Goodfellow, Yoshua Bengio, *Deep Learning*. MIT Press, 2016.

[6] F. H. J. C. A. C. A. A. A. A. A. T. J. T. Z. W. W. S. Christian Ledig, Lucas Theis, "Photo-realistic single image super-resolution using a generative adversarial network," *IEEE Computer Vision Foundation*, pp. 4681–4690, 2016.

[7] GeeksforGeeks. Super resolution gan (sran). Accessed: 2025-10-21. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/super-resolution-gan-srgan/

[8] S. W. J. G. Y. L. C. D. Y. Q. C. C. L. Xintao Wang, Ke Yu, "Esrgan: Enhanced super-resolution generative adversarial networks," *European Conference on Computer Vision (ECCV) Workshops*, vol. 11133, pp. 63–79, 2018.

[9] V. Hoàng. Tìm hiểu về yolo trong bài toán real-time object detection. Accessed: 2025-10-21. [Online]. Available: https://viblo.asia/p/tim-hieu-ve-yolo-trong-bai-toan-real-time-object-detection-yMnKMdvr57P