**­­­­­ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**TRƯỜNG CNTT&TT**



**ĐỒ ÁN MÔN HỌC**

**HỌC SÂU VÀ ỨNG DỤNG**

**Tên đề tài:**

**Phát hiện Vùng Quan Tâm trong Hình Ảnh Y Tế với LSM**

|  |  |  |
| --- | --- | --- |
| **Nhóm sinh viên thực hiện:** | Ngô Trí Cảnh  Vũ Đình Cường | 20220015  20224942 |
| **Lớp:** | 159467 | |
| **Giáo viên hướng dẫn:** | PGS. TS. Nguyễn Hồng Quang | |
| **HÀ NỘI, 05/2025** | | |

Mục lục

[Chương 1. Tổng quan đề tài 6](#_Toc198497812)

[1.1. Đặt vấn đề 6](#_Toc198497813)

[1.2. Các nghiên cứu liên quan 7](#_Toc198497814)

[1.3. Bố cục đồ án 8](#_Toc198497815)

[Chương 2. Dữ liệu 10](#_Toc198497816)

[2.1. Giới thiệu chung về tập dữ liệu 10](#_Toc198497817)

[2.2. Mô tả chi tiết tập dữ liệu 10](#_Toc198497818)

[2.3. Tiền xử lý dữ liệu 12](#_Toc198497819)

[2.3.1. Định dạng lại nhãn theo định dạng YOLO 12](#_Toc198497820)

[2.3.2. Làm giàu dữ liệu 13](#_Toc198497821)

[2.4. Kết quả thu được 13](#_Toc198497822)

[Chương 3. Các phương pháp đánh giá 14](#_Toc198497823)

[3.1. mAP-50 (Mean Average Precision at IoU=0.5) 14](#_Toc198497824)

[3.2. mAP-95 (COCO-style mAP) 14](#_Toc198497825)

[3.3. Precision (P) 14](#_Toc198497826)

[3.4. Recall (R) 14](#_Toc198497827)

[3.5. GFLOPs (Giga Floating-point Operations) 15](#_Toc198497828)

[3.6. Parameters (Số lượng tham số) 15](#_Toc198497829)

[Chương 4. LSM YOLO 16](#_Toc198497830)

[4.1. Lightweight Adaptive Extraction (LAE) 16](#_Toc198497831)

[4.2. Multipath Shunt Feature Matching (MSFM) 18](#_Toc198497832)

[4.2.1. Đầu vào và Luồng rẽ nhánh 18](#_Toc198497833)

[4.2.2. Khối MatchNeck 18](#_Toc198497834)

[4.2.3. Hai loại MSFM khi sử dụng: 19](#_Toc198497835)

[4.2.4. Kết quả 19](#_Toc198497836)

[4.3. Hàm Loss tối ưu 20](#_Toc198497837)

[4.4. RFAConv & PA-FPN 20](#_Toc198497838)

[4.4.1. RFAConv [7] 20](#_Toc198497839)

[4.4.2. PA-FPN [8] 21](#_Toc198497840)

[4.5. Kết quả 21](#_Toc198497841)

[Chương 5. Phương pháp đề xuất 23](#_Toc198497842)

[5.1. Giới thiệu về kỹ thuật Knowledge Distillation 23](#_Toc198497843)

[5.1.1. Các hình thức Knowledge Distillation 23](#_Toc198497844)

[5.1.2. Thuật toán Knowledge Distillation 27](#_Toc198497845)

[5.2. Áp dụng kỹ thuật Knowledge Distillation với LSM-YOLO 29](#_Toc198497846)

[5.2.1. Mô hình teacher 29](#_Toc198497847)

[5.2.2. Knowledge 30](#_Toc198497848)

[5.2.3. Kiến trúc teacher-student 31](#_Toc198497849)

[Chương 6. Kết quả thực nghiệm 32](#_Toc198497850)

[6.1. Các thông số cho quá trình chưng cất tri thức với LSM-YOLO 32](#_Toc198497851)

[6.2. Kết quả thức nghiệm. 32](#_Toc198497852)

[Tài liệu tham khảo 34](#_Toc198497853)

# Danh mục hình ảnh

[Hình 1. Trước và sau khi xử lý 11](#_Toc198739777)

[Hình 2. Phân bố nhãn trên từng tập dữ liệu 11](#_Toc198739778)

[Hình 3. Ảnh mẫu MRI 11](#_Toc198739779)

[Hình 4. Ảnh mẫu MRI với ROI 12](#_Toc198739780)

[Hình 5. Cấu trúc thư mục dữ liệu huấn luyện 12](#_Toc198739781)

[Hình 6. Định dạng lại nhãn 13](#_Toc198739782)

[Hình 7. Kiến trúc mạng LSM-YOLO 16](#_Toc198739783)

[Hình 8. Lightweight Adaptive Extraction 17](#_Toc198739784)

[Hình 9. Multipath Shunt Feature Matching (MSFM) 18](#_Toc198739785)

[Hình 10. Knowledge Distillation 23](#_Toc198739786)

[Hình 11. Các loại tri thức khác nhau trong mạng nơ-ron 24](#_Toc198739787)

[Hình 12. Response-based knowledge distillation 24](#_Toc198739788)

[Hình 13, Feature-based knowledge distillation 25](#_Toc198739789)

[Hình 14. Relation-based knowledge distillation 26](#_Toc198739790)

[Hình 15. Types of knowledge distillation training schemes 27](#_Toc198739791)

[Hình 16. Multi-teacher distillation 28](#_Toc198739792)

[Hình 17. Cross-modal distillation 28](#_Toc198739793)

[Hình 18. Data-Free Distillation 29](#_Toc198739794)

# Chương 1. Tổng quan đề tài

## 1.1. Đặt vấn đề

Trong lĩnh vực y tế hiện đại, ảnh y tế như X-quang, CT (Chụp cắt lớp), MRI (Cộng hưởng từ) và siêu âm đóng vai trò thiết yếu trong việc hỗ trợ chẩn đoán và điều trị bệnh. Những hình ảnh này cung cấp thông tin chi tiết về cấu trúc bên trong cơ thể, giúp bác sĩ nhận diện các bất thường như khối u, vết gãy xương, hoặc tổn thương mô. Tuy nhiên, việc phân tích thủ công các ảnh y tế đòi hỏi thời gian, kinh nghiệm chuyên môn cao và dễ xảy ra sai sót do sự khác biệt giữa các nhà phân tích (inter-observer variability). Trong bối cảnh khối lượng dữ liệu y tế ngày càng tăng và yêu cầu chẩn đoán nhanh, đặc biệt trong các tình huống cấp cứu, việc tự động hóa quá trình phân tích ảnh y tế trở thành một nhu cầu cấp thiết.

Một nhiệm vụ quan trọng trong phân tích ảnh y tế là phát hiện vùng quan tâm (Region of Interest - ROI), tức là các khu vực cụ thể trong ảnh chứa đựng thông tin chẩn đoán quan trọng, như khối u não trong ảnh MRI hay vết nứt xương trong ảnh X-quang. Phát hiện chính xác ROI không chỉ giúp nâng cao độ chính xác của chẩn đoán mà còn hỗ trợ bác sĩ đưa ra quyết định nhanh chóng, cải thiện kết quả điều trị và giảm chi phí y tế. Tuy nhiên, bài toán này đối mặt với nhiều thách thức lớn:

* **Độ phức tạp của ảnh y tế**: Ảnh y tế thường có nhiễu, độ tương phản thấp và sự đa dạng về cấu trúc giải phẫu, khiến việc xác định ROI trở nên khó khăn.
* **Biến thiên chất lượng ảnh**: Sự khác biệt về thiết bị quét, điều kiện chụp và đặc điểm bệnh nhân (tuổi, giới tính, bệnh lý nền) làm ảnh hưởng đến tính nhất quán của dữ liệu.
* **Yêu cầu độ chính xác cao**: Sai lầm trong việc phát hiện hoặc bỏ sót ROI có thể dẫn đến chẩn đoán sai, gây hậu quả nghiêm trọng cho sức khỏe bệnh nhân.
* **Khối lượng dữ liệu lớn**: Các cơ sở y tế phải xử lý hàng ngàn ảnh mỗi ngày, đòi hỏi phương pháp phân tích nhanh chóng và hiệu quả.

Để giải quyết các thách thức này, các phương pháp học sâu, đặc biệt là thuật toán **YOLO (You Only Look Once)**, đã được ứng dụng để tự động hóa việc phát hiện ROI. YOLO nổi bật với khả năng xử lý ảnh trong thời gian thực, cung cấp các vùng quan tâm dưới dạng bounding boxes với tốc độ và độ chính xác cao, phù hợp cho các ứng dụng chẩn đoán nhanh. Tuy nhiên, việc áp dụng YOLO vào ảnh y tế vẫn cần vượt qua các rào cản như nhiễu, yêu cầu dữ liệu huấn luyện lớn và đảm bảo độ chính xác lâm sàng.

Báo cáo này tập trung vào bài toán phát hiện vùng quan tâm trong ảnh y tế, đặc biệt là ảnh CT của não, với trọng tâm là ứng dụng của YOLO. Mục tiêu là đánh giá tiềm năng của YOLO trong việc hỗ trợ chẩn đoán, phân tích các thách thức kỹ thuật và đề xuất hướng cải tiến để nâng cao hiệu quả ứng dụng trong thực tiễn y tế. Việc phát triển các công cụ như YOLO không chỉ giảm gánh nặng cho các chuyên gia y tế mà còn góp phần xây dựng các hệ thống y tế thông minh, mang lại lợi ích thiết thực cho bệnh nhân và cộng đồng.

## 1.2. Các nghiên cứu liên quan

Từ năm 2020, với sự phát triển của Computer Vision, đặc biệt là các phương pháp học sâu (Deep Learning), đã được ứng dụng rộng rãi để tự động hóa việc phát hiện vùng quan tâm ROI. Dưới đây là một số nghiên cứu nổi bật.

* **Region of interest (ROI) selection using vision transformer for automatic analysis using whole slide images** [[1]](#ref2), tác giả đề xuất một phương pháp hoàn toàn dựa trên Vision Transformer (ViT) nhằm tự động phát hiện ROI trên ảnh mô toàn bộ (WSI), thay thế bước chọn ROI thủ công thường dùng. Mô hình ViT được huấn luyện trên các patch WSI ở độ phóng đại 10× và 20×, tận dụng cơ chế self‑attention để học biểu diễn toàn cục của khung mô và từ đó xác định các patch có chứa thông tin chẩn đoán quan trọng . Phương pháp này đạt độ chính xác 99% ở 20x và 97% ở 10x cho nhiệm vụ phát hiện ROI, đồng thời rút ngắn thời gian xử lý xuống dưới 15 giây so với 30–60 phút của thủ tục thủ công nhờ chỉ phân tích những khu vực đã chọn trước.
* **Deep Learning for Medical Image-Based Cancer Diagnosis** [[2]](#ref3). Nghiên cứu cung cấp một cái nhìn tổng quan về các kiến trúc Deep Learning - bao gồm Transfer Learning, Ensemble Learning, Graph Neural Network và đặc biệt là Vision Transformer với các biến thể như ASI‑DBNet, Multi‑scale ViT và SI‑ViT - cho các nhiệm vụ phân loại, phát hiện và phân đoạn ROI trên X‑quang, CT, MRI và ảnh mô học . Trong đó, cơ chế self‑attention của ViT được nhấn mạnh vì khả năng nắm bắt mối liên hệ toàn cục giữa các vùng, giúp cải thiện hiệu suất đặc biệt khi dữ liệu hạn chế; đồng thời tác giả cũng đề xuất áp dụng các kỹ thuật chống overfitting như batch normalization, dropout và data augmentation để tăng độ ổn định của mô hình . Kết quả tổng hợp cho thấy các phương pháp này không chỉ nâng cao độ chính xác trong chẩn đoán mà còn có tiềm năng tích hợp vào hệ thống lâm sàng nhờ tính linh hoạt và khả năng mở rộng
* **Deep Learning-Based Object Detection Strategies for Disease Detection and Localization in Chest X-Ray Images** [[3]](#ref1). Tác giả đã phát triển và so sánh các chiến lược object detection trên ảnh CXR, bao gồm Faster R‑CNN cải tiến, DETR và Meta‑DETR, dùng dữ liệu chú thích bounding box từ Bệnh viện E‑Da. Trong đó, Faster R‑CNN được nâng cao bằng cách bổ sung hàm mất mát tương phản (“contrastive proposal encoding loss”) để cải thiện chất lượng các đề xuất vùng, còn Meta‑DETR tích hợp mô‑đun tổng hợp tương quan (correlation synthesis module) nhằm nâng cao khả năng phân biệt giữa các dấu hiệu bệnh lý. Để giải quyết bài toán dữ liệu hiếm, nhóm tác giả còn áp dụng kỹ thuật Few‑Shot Object Detection (1‑shot đến 10‑shot) và điều chỉnh tỉ lệ ảnh nền/bệnh lý trong tập huấn luyện, kết quả Faster R‑CNN cải tiến vẫn cho mAP cao hơn so với các mô hình Transformer, đồng thời đề xuất thêm bước phân loại nhị phân (binary classification) sau detection để tăng độ chính xác cuối cùng.

Kết quả từ các nghiên cứu vừa rồi cho thấy việc kết hợp cơ chế self‑attention của Transformer với các kiến trúc CNN cải tiến giúp phát hiện ROI trên ảnh y tế vừa chính xác vừa nhanh chóng. Tuy nhiên, nhiều mô hình vẫn có kích thước lớn và khó triển khai trong môi trường thời gian thực tại bệnh viện. LSM‑YOLO (Lightweight Shunt Matching‑YOLO) được đề xuất để khắc phục hạn chế này bằng kiến trúc gọn nhẹ nhưng vẫn duy trì hiệu suất cao. Thành phần LAE (Lightweight Adaptive Extraction) của LSM‑YOLO cho phép trích xuất đặc trưng đa quy mô với chi phí tham số thấp, trong khi MSFM (Multipath Shunt Feature Matching) tập trung tinh chỉnh và hợp nhất thông tin từ các cấp độ feature map khác nhau. Nhờ kết hợp hai kỹ thuật này, LSM‑YOLO đạt được độ chính xác cạnh tranh trong khi giảm đáng kể độ phức tạp tính toán. Tiếp theo, chúng ta sẽ phân tích chi tiết cấu trúc của LSM‑YOLO và so sánh hiệu quả thực nghiệm với các phương pháp ROI detection hiện có.

## 1.3. Bố cục đồ án

Phần còn lại của báo cáo sẽ được tổ chức như sau:

* Chương 2 cung cấp thông tin về bộ dữ liệu Br35H được sử dụng trong qua trình huấn luyện mô hình
* Chương 3 trình bày các tiêu chí đánh giá chính được sử dụng để đánh giá độ tốt của mô hình cũng như là số liệu để tham chiếu đến các nghiên cứu liên quan
* Chương 4 sẽ tập trung cung cấp, giới thiệu về LSM-YOLO, một thuật toán đề xuất giúp tăng hiệu năng trong quá trình phát hiện vùng quan tâm
* Chương 5 trình bày những đóng góp chính của chúng em vào việc việc cải thiện hiệu suất của mô hình LSM-YOLO thông qua kỹ thuật chưng cất tri thức (Knowledge Distillation)
* Chương 6 sẽ đánh giá và đưa ra kết quả cuối cùng của phương pháp đề suất với các giải thuật nền tảng trước đó.

# Chương 2. Dữ liệu

## 2.1. Giới thiệu chung về tập dữ liệu

Dữ liệu đóng vai trò cốt lõi và không thể thiếu trong việc huấn luyện và đánh giá các mô hình trí tuệ nhân tạo, đặc biệt là mô hình LSM-YOLO, được thiết kế để tự động phát hiện vùng quan tâm (ROI) trong ảnh y tế. Nghiên cứu này sử dụng tập dữ liệu công khai Br35H[[4]](#ref4), tải từ [Kaggle](https://www.kaggle.com/), ban đầu gồm 501 ảnh huấn luyện, 200 ảnh đánh giá, và 101 ảnh kiểm tra, với các ảnh MRI của não được chú thích để hỗ trợ phát hiện khối u não. Để nâng cao chất lượng và phù hợp với mục tiêu nghiên cứu, tập dữ liệu đã được tiền xử lý, bổ sung thêm dữ liệu, và định dạng lại nhãn theo phương pháp riêng (chi tiết trình bày ở mục [2.3](#_2.3._Tiền_xử)). Br35H được lựa chọn nhờ chất lượng hình ảnh cao, nhãn chú thích đáng tin cậy, và tính đại diện cho chẩn đoán ung thư não trong thực tế lâm sàng. Sau xử lý, tập dữ liệu đảm bảo đáp ứng yêu cầu của LSM-YOLO, nâng cao khả năng tổng quát hóa mô hình. Phần tiếp theo sẽ trình bày chi tiết đặc điểm, nguồn gốc, và quy trình tiền xử lý của tập dữ liệu Br35H.

## 2.2. Mô tả chi tiết tập dữ liệu

Tập dữ liệu Br35H[[4]](#ref4)là nguồn dữ liệu chính để huấn luyện và đánh giá mô và cải tiến mô hình LSM-YOLO trong nhiệm vụ phát hiện vùng quan tâm (ROI) liên quan đến khối u não trong ảnh y tế. Ban đầu, Br35H, tải từ [Kaggle](https://www.kaggle.com/), bao gồm 802 ảnh MRI của não, được chia thành 501 ảnh huấn luyện, 200 ảnh xác thực, và 101 ảnh kiểm tra. Các ảnh này có độ phân giải 256x256, định dạng PNG, và được chú thích với đánh nhãn phân đoạn ngữ nghĩa (polygon annotations) để xác định vị trí và hình dạng khối u não hỗ trợ chẩn đoán ung thư não trong thực tế lâm sàng. Sau khi thực hiện tiền xử lý, bổ sung dữ liệu, và định dạng lại nhãn, tập dữ liệu đã được tối ưu hóa để phù hợp với yêu cầu của LSM-YOLO. Các thông số sẽ được tringh bày trong bảng 2.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Trạng thái | Train | Validation | Test | Nhãn |
| Ban đầu | 501 | 200 | 101 | Polygon annotations |
| Đã xử lý | 1251 | 376 | 376 | Bounding Boxs |

Bảng 1. Số lượng ảnh và nhãn trong tập Br35H trước và sau xử lý

A graph with blue and orange bars

Description automatically generated

Hình . Trước và sau khi xử lý

A graph of different colored bars

Description automatically generated

Hình . Phân bố nhãn trên từng tập dữ liệu

Để minh họa chất lượng hình ảnh và kiểu chú thích của bộ dữ liệu Br35H được sử dụng, Hình 3, hình 4 sẽ trình bày một số hình ảnh mẫu với các khối u được khoanh vùng với các hộp định vị.

A close-up of a brain scan

Description automatically generatedA close-up of a brain scan

Description automatically generatedA close-up of a brain scan

Description automatically generatedA close-up of a brain scan

Description automatically generated

Hình . Ảnh mẫu MRI

A close-up of a brain scan

Description automatically generatedA close-up of a brain scan

Description automatically generatedA close-up of a brain scan

Description automatically generatedA close-up of a brain scan

Description automatically generated

Hình . Ảnh mẫu MRI với ROI

Cấu trúc thư mục dữ liệu được tổ chức theo định dạng YOLO: train, val, test.

A screen shot of a computer program

Description automatically generated

Hình . Cấu trúc thư mục dữ liệu huấn luyện

## 2.3. Tiền xử lý dữ liệu

Trong phần này, chúng tôi sẽ trình bày chi tiết các bước tiền xử lý dữ liệu. Như đã đề cập ở trên, bộ dữ liệu được tải về từ Kaggle, bao gồm 801 hình ảnh, dưới dạng định dạng phân đoạn ngữ nghĩa. Các bước xử lý chi tiết sẽ được trình bày như sau:

### 2.3.1. Định dạng lại nhãn theo định dạng YOLO

Để phục vụ quá trình huấn luyện mô hình YOLO, dữ liệu cần được chuẩn hóa theo định dạng phù hợp. Ban đầu, các ảnh trong bộ dữ liệu được gán nhãn bằng tệp JSON theo định dạng VIA (VGG Image Annotator).

Tuy nhiên, định dạng này không tương thích trực tiếp với mô hình YOLO. Do đó, cần thực hiện chuyển đổi dữ liệu bằng cách:

* **Tính toán bounding box** từ các điểm đa giác (polygon) được cung cấp.
* **Xác định nhãn lớp** (class) dựa trên thông tin trong trường region\_attributes.
* **Xuất thông tin sang định dạng YOLO**, trong đó mỗi dòng trong tệp .txt tương ứng với một đối tượng, có cấu trúc:

<class\_id> <x\_center> <y\_center> <width> <height>

A diagram with a black line

Description automatically generated with medium confidence

Hình . Định dạng lại nhãn

### 2.3.2. Làm giàu dữ liệu

Ban đầu, bộ dữ liệu gốc chỉ bao gồm 802 ảnh, được chia đều cho ba tập: huấn luyện (train), xác thực (validation), và kiểm tra (test). Tuy nhiên, số lượng này là chưa đủ để mô hình học sâu như YOLO đạt được hiệu quả tối ưu trong quá trình huấn luyện.

Do đó, để tăng tính đa dạng và độ phủ dữ liệu, chúng tôi đã tiến hành làm giàu tập dữ liệu bằng cách bổ sung thủ công một số ảnh mới. Việc thêm dữ liệu được thực hiện một cách có kiểm soát nhằm đảm bảo giữ nguyên tỷ lệ phân chia giữa ba tập dữ liệu (train, val, test), từ đó tránh gây lệch phân phối hoặc ảnh hưởng đến quá trình đánh giá hiệu suất của mô hình.

## 2.4. Kết quả thu được

Sau khi hoàn tất quá trình tiền xử lý dữ liệu, bao gồm chuyển đổi định dạng nhãn từ **VIA JSON** sang định dạng chuẩn của YsOLO, cùng với việc làm giàu bộ dữ liệu, chúng tôi thu được các kết quả như sau:

* Tổng số ảnh sau khi mở rộng: 2003 ảnh (tăng từ 802 ảnh ban đầu).
* Tất cả các nhãn đã được chuyển đổi thành định dạng .txt tương ứng với mỗi ảnh, tuân theo cấu trúc định dạng của YOLO.
* Bounding box được tính toán tự động từ các polygon
* Tên lớp (class name) được ánh xạ chính xác sang các mã số lớp (class ID) dựa trên thuộc tính <region\_attributes>

Kết quả này đảm bảo bộ dữ liệu đã hoàn toàn sẵn sàng để đưa vào huấn luyện với mô hình YOLO.

# Chương 3. Các phương pháp đánh giá

Trong bài toán phát hiện vùng quan tâm (Region of Interest - ROI) từ ảnh y tế, việc lựa chọn các chỉ số đánh giá phù hợp đóng vai trò quan trọng trong việc phản ánh chính xác hiệu suất của mô hình. Các chỉ số không chỉ đo lường độ chính xác về mặt kỹ thuật mà còn thể hiện mức độ phù hợp để ứng dụng trong môi trường lâm sàng thực tế, nơi các sai sót có thể dẫn đến hậu quả nghiêm trọng. Trong báo cáo này, các chỉ số sau được lựa chọn để đánh giá mô hình.

## 3.1. mAP-50 (Mean Average Precision at IoU=0.5)

Đây là một trong những chỉ số quan trọng và được sử dụng rộng rãi trong các bài toán phát hiện đối tượng. mAP-50 phản ánh trung bình độ chính xác (Average Precision - AP) khi ngưỡng chồng lấn giữa vùng dự đoán và vùng thực (Intersection over Union - IoU) lớn hơn hoặc bằng 0.5. Điều này có nghĩa là mô hình chỉ cần dự đoán trúng vị trí và loại đối tượng trong khoảng sai số cho phép nhất định. mAP@0.5 thường được dùng để đánh giá mô hình trong môi trường ứng dụng thực tế, nơi mức độ chính xác tuyệt đối là không cần thiết, nhưng độ tin cậy cao là bắt buộc. Giá trị mAP cao đồng nghĩa với khả năng phát hiện đối tượng tốt của mô hình.

## 3.2. mAP-95 (COCO-style mAP)

Khác với mAP-50, chỉ số mAP-95 được tính trung bình qua nhiều ngưỡng IoU từ 0.5 đến 0.95 với bước nhảy 0.05. Đây là chỉ số được sử dụng trong bộ dữ liệu COCO để đánh giá độ chính xác toàn diện của mô hình ở các mức độ khó khác nhau. Một mô hình có mAP@0.5 cao nhưng mAP@0.5:0.95 thấp cho thấy khả năng phát hiện không ổn định và dễ mất chính xác khi yêu cầu khắt khe hơn. Do đó, mAP@0.5:0.95 được xem là thước đo "nghiêm ngặt" để kiểm tra tính tổng quát và độ bền vững của mô hình trong điều kiện thay đổi.

## 3.3. Precision (P)

Precision đo lường tỷ lệ các dự đoán đúng trong tất cả các dự đoán mà mô hình đã đưa ra. Nó đặc biệt quan trọng trong các hệ thống y tế vì một mô hình có precision thấp có thể đưa ra nhiều cảnh báo sai (false positives), khiến bác sĩ phải kiểm tra lại nhiều lần, làm tăng khối lượng công việc và khả năng dẫn đến sai sót. Một mô hình có precision cao sẽ đảm bảo rằng khi mô hình đưa ra cảnh báo, nó thực sự đáng tin cậy.

Công thức: .

## 3.4. Recall (R)

Recall phản ánh khả năng phát hiện đầy đủ các đối tượng thực sự tồn tại trong ảnh. Đây là chỉ số đặc biệt quan trọng trong các ứng dụng y tế, nơi việc bỏ sót một vùng bất thường (False Negative) có thể dẫn đến chẩn đoán sai và ảnh hưởng nghiêm trọng đến sức khỏe người bệnh. Mô hình có recall cao đảm bảo phát hiện hầu hết các vùng nghi ngờ, giúp bác sĩ không bỏ sót trường hợp quan trọng.

Công thức:

## 3.5. GFLOPs (Giga Floating-point Operations)

GFLOPs là thước đo số lượng phép tính dấu phẩy động (floating-point operations) cần thiết để mô hình xử lý một ảnh đầu vào. GFLOPs phản ánh độ phức tạp tính toán của mô hình và ảnh hưởng đến tốc độ suy luận, tiêu thụ tài nguyên phần cứng và mức độ phù hợp triển khai thực tế. Trong các ứng dụng y tế, đặc biệt tại các cơ sở tuyến dưới hoặc thiết bị y tế cầm tay, mô hình nhẹ với GFLOPs thấp sẽ có lợi thế lớn về tốc độ và khả năng hoạt động ổn định trong điều kiện hạn chế.

## 3.6. Parameters (Số lượng tham số)

GFLOPs là thước đo số lượng phép tính dấu phẩy động (floating-point operations) cần thiết để mô hình xử lý một ảnh đầu vào. GFLOPs phản ánh độ phức tạp tính toán của mô hình và ảnh hưởng đến tốc độ suy luận, tiêu thụ tài nguyên phần cứng và mức độ phù hợp triển khai thực tế. Trong các ứng dụng y tế, đặc biệt tại các cơ sở tuyến dưới hoặc thiết bị y tế cầm tay, mô hình nhẹ với GFLOPs thấp sẽ có lợi thế lớn về tốc độ và khả năng hoạt động ổn định trong điều kiện hạn chế.

Việc sử dụng đồng thời các chỉ số mAP@0.5, mAP@0.5:0.95, Precision, Recall, GFLOPs và Parameters cho phép đánh giá mô hình một cách toàn diện về cả độ chính xác, tính hiệu quả và khả năng triển khai thực tế. Trong nghiên cứu này, các chỉ số trên sẽ được sử dụng để so sánh LSM-YOLO với các mô hình khác như YOLOv5, Faster R-CNN và các mô hình dựa trên Transformer, từ đó làm rõ điểm mạnh và tính ứng dụng của LSM-YOLO trong phát hiện ROI trên ảnh y tế.

# Chương 4. LSM YOLO

Trong các nghiên cứu gần đây ([mục 1.2.](#_1.2._Các_nghiên)), chúng ta đã đề cập khá nhiều phương pháp giúp phát hiện vùng quan tâm trong ảnh y tế. Nhung một vấn đề gặp phải trong các phương pháp trên, số lượng tham số và khối lượng tính toán khá lớn trong quá trình huấn luyện cững như dự đoán. Một nhu cầu cần thiết về một mô hình mới với số lượng tham số ít hơn, khôi lượng tính toán ít hơn nhưng vẫn đạt được hiệu suất cao như các mô hình tước đó. Đó là tiền đề cho sự ra đời của Lightweight Shunt Matching-YOLO (LSM-YOLO) [[5]](#ref5), với Lightweight Adaptive Extraction (LAE) và Multipath Shunt Feature Matching (MSFM).

Kiến trúc mạng LSM-YOLO được thể hiện trong Hình 7

A diagram of a computer system

Description automatically generated

Hình . Kiến trúc mạng LSM-YOLO

## 4.1. Lightweight Adaptive Extraction (LAE)

Trong các phiên bản YOLO truyền thống, các phương pháp Downsampling thường sử dụng các lớp tích chập với stride lớn hoặc các lớp pooling. Điều này vô tình làm mất thông tin của các pixel ở cạnh, ở góc hoặc các vật thể nhỏ, đây là một vấn đề có tác động rất lớn, đặc biệt là trong ảnh y tế. Mặt khác các lớp tích chập truyền thống cũng tốn nhiều tìa nguyên tính toán. Sự ra đời của LAE khắc phục các nhược điểm trên của các phương pháp Downsampling truyền thống. LAE “thông minh” và “gọn lẹ” hơn các phương pháp truyền thống trong việc giảm chiều FeatureMap đầu vào. Hình 8 thể hiện cấu trúc của LAE

A diagram of a diagram of a variety of blocks

Description automatically generated with medium confidence

Hình . Lightweight Adaptive Extraction

LAE thực hiện qua hai nhánh song sonh với nhau, kết quả ta thu được một Feature Map có chiều cao và rộng giảm đi một nửa:

* Lightweight Extraction: thay vì dùng phương pháp tích chập truyền thống, LAE sử dụng kỹ thuật Group Convolution. Với kỹ thuật này, ta thu được một Feature Map với các đặc trung đã được “cô đặc” hơn, đồng thời cũng giúp giảm khối lượng tính toán và khối lượng tham số thông qua kỹ thuật tích chập nhóm.
* Adaptive Extraction: với nhánh này, Feature map đầu vào đi qua 1 lớp Average Pooling, 1 lớp tích chập, 1 lớp Softmax như một mạng tích chập thông thường. Nhánh này giúp ta tính trọng số (hay mức độ quan trong) của từng vùng khác nhau trong bản đồ đặc trưng được trích suất ra từ nhánh LAE.
* Dimension Mapping: ở cả hai nhánh, trước khi được kết hợp với nhau, đều phải đi qua một bước gọi là “ánh xạ chiều”. Dimension Mapping tái cáu trúc tensor đặc trưng từ 4 chiều sang 5 chiều, bằng cách thêm 1 đặc trung nữa là yếu tố lấy mẫu “n”. Bằng cách làm việc với n chiều và n trọng số tương ứng, mô hình có thể học được từ pixel gốc một cách linh hoạt và thông minh hớn, thay vid chỉ làm việc với 1 phép toán cố định như AveragePooling hay MaxPooling.
* Kết hợp kết quả 2 nhánh:
  + Nhân trọng số từ nhánh AE với phần tử tương ứng trong nhánh LE: giúp mô hình học được những thông tin quan trọng và bỏ qua các thông tin ít quan trọng hơn
  + SUM: Cộng tổng các phần tử ở cùng một chiều n sau khi nhân để tạo ra Feature Map cuối cùng. Đây cũng chính là đầu ra của LAE.

Như vậy, với các kỹ thuật trên, ta có thể trích suất thông tin một cách chính xác và đầy đủ hơn nhưng vẫn giảm được đáng kể số lượng tham số được sử dụng.

## 4.2. Multipath Shunt Feature Matching (MSFM)

**Multipath Shunt Feature Matching (MSFM)** là một mô-đun cốt lõi trong mô hình **LSM-YOLO**, được thiết kế để tối ưu hóa việc kết hợp đặc trưng đa tỷ lệ (multi-scale feature fusion) trong phát hiện vùng quan tâm (ROI) trên ảnh y tế. MSFM giải quyết các thách thức đặc thù của hình ảnh y khoa, như kích thước nhỏ của ROI, hiện tượng chồng lấn, và mối quan hệ phức tạp giữa ROI với vùng lân cận, đồng thời đảm bảo tính nhẹ (lightweight) để triển khai trên các thiết bị y tế.

Đây là một module khá phức tạp nhưng hiệu quả.

A diagram of a computer

Description automatically generated

Hình . Multipath Shunt Feature Matching (MSFM)

### 4.2.1. Đầu vào và Luồng rẽ nhánh

Đầu vào của MFSM là một tensor đặc trưng . Một toán tử được sử dụng để chia thông tin đầu vào, giữ lại các thông tin gốc quan trọng để sử dụng trong kết nối dư (Residual) sau này (). MFSM sẽ trích xuất thông tin từ ba chiều: chiều cao (height), chiều rộng (weight) và kênh (channel).

### 4.2.2. Khối MatchNeck

Đây là một khối vô cùng quan trọng, có thể được coi là trái tim của MSFM.

* Trích suất thông tin từ không gian và kênh (IE). Khối này sử dụng các phép average pooling () and global average pooling () để trích suất thông tin từ nhiều chiều khác nhau
  + : thông tin chiều cao.
  + : thông tin chiều chiều rộng.
  + : thông tin theo kênh.
* Các đặc trưng trên được căn chỉnh với kết nối dư để tạo ra một đầu ra tạm thời:

.

Đây cũng chính là đâu ra của khối IE.

* Xử lý thông tin không gian (Spatail Information)
  + Thông tin về chiều cao (height), và chiều rộng (weight) tiếp tục được kết nối (Concat) và chia lại:
  + Ngoài ra các trọng số không gian cúng được trích xuất:
* Xử lý thông tin kênh (Channel Information): điều chỉnh bằng cách nhân với thông tin không gian:
* Kết quả đầu ra của khối MSFM thu được bằng cách kết hợp các thông tin về các chiều không gian và kết nối dư:

### 4.2.3. Hai loại MSFM khi sử dụng:

* Với kết nối dư (residual connections): Được sử dụng trong backbone để trích xuất đặc trưng chính, giúp giảm thiểu vấn đề gradient biến mất và tăng tốc hội tụ.
* Không có kết nối dư: Được sử dụng trong head, nơi mô hình tập trung vào phân tích đặc trưng để dự đoán đối tượng, do đó không cần kết nối dư.

### 4.2.4. Kết quả

Như vậy ta có thế thấy, MSFM là khối khá phức tạp nhưng mang lại hiệu quả cao. Việc trích xuất thông tin từ các kênh không gian giúp cải thiện việc phát hiện vùng quan tâm ROI trong ảnh y tế. Việc này mang lại 2 lợi ích lớn:

* Học được mối quan hệ giữa vùng ROI với các vùng lân cận: MSFM trích xuất thông tin từ chiều cao và chiều rộng giúp mô hình nhận diện các tương quan không gian, như hiện tượng xâm lấn hoặc di căn của khối u, bằng cách tập trung vào sự tương phản giữa ROI và mô xung quanh.
* **Phân biệt các vùng nhỏ và chồng lấn**: Bằng cách kết hợp thông tin không gian với thông tin kênh () qua cơ chế phân luồng (shunt), MSFM hỗ trợ phát hiện các ROI nhỏ và chồng lấn.

## 4.3. Hàm Loss tối ưu

Trong mô hình LSM-YOLO, hàm loss tối ưu hóa được thiết kế để cân bằng giữa nhiệm vụ **phân loại** (classification) và **hồi quy hộp giới hạn** (bounding box regression), giúp cải thiện độ chính xác trong phát hiện vùng quan tâm (ROI) trên ảnh y tế.

Trong đó:

Hàm loss gồm 3 thành phần:

* Binary Cross-Entropy Loss: được sử dụng cho nhánh phân loại
* Distribution Focal Loss: được sử dụng cho nhánh hồi quy để dự đoán phân phối xác suất của các hộp giới hạn (Biunding Boxes)
* SCYLLA-IoU Loss (SIoU) [[6]](#ref6): là một biến thể của IoU Loss, được thiết kế để cải thiện hồi quy hộp giới hạn bằng cách xem xét không chỉ mức độ giao nhau (intersection over union) mà còn các yếu tố như khoảng cách, hình dạng, và góc giữa các hộp giới hạn dự đoán và thực tế.

## 4.4. RFAConv & PA-FPN

Ngoài 3 kỹ thuật chính được tác giả đề xuất, họ còn sử dụng 2 kỹ thuật RFAConv & PA-FPN để cải thiện hiệu suất của mô hình.

### 4.4.1. RFAConv [[7]](#ref7)

Đây là một phép toán tích chập cải tiến, kết hợp cơ chế chú ý không gian để chia sẻ tham số nhân tích chập (convolutional kernel parameters). Điều này giúp mô hình tập trung vào các vùng quan trọng trong ảnh và học được sự tương đồng cũng như khác biệt giữa ROI và các vùng lân cận.

Cơ chế:

* RFAConv áp dụng chú ý không gian để điều chỉnh trọng số của các vùng trong trường tiếp nhận (receptive field), tăng cường khả năng nhận diện các đặc điểm quan trọng như ranh giới hoặc cấu trúc của ROI.
* Thay vì sử dụng tích chập truyền thống với số lượng tham số lớn, RFAConv giảm chi phí tính toán bằng cách tái sử dụng các tham số nhân tích chập, duy trì tính nhẹ (lightweight) của mô hình.

Lợi ích:

* **Mở rộng trường tiếp nhận**: Tăng khả năng nắm bắt thông tin ngữ cảnh toàn cục, phù hợp với các ROI phức tạp như khối u trong ảnh y tế.
* **Cải thiện phát hiện ROI tinh vi**: Tập trung vào ranh giới và đặc điểm không gian, nâng cao độ chính xác cho ROI nhỏ
* **Thiết kế nhẹ**: Giảm chi phí tính toán, hỗ trợ triển khai trên thiết bị y tế hạn chế tài nguyên.

### 4.4.2. PA-FPN[[8]](#ref8)

Đây là một cấu trúc mạng kim tự tháp đặc trưng được mở rộng với bốn đầu ra (detection heads) trong LSM-YOLO, tích hợp đặc trưng đa tỷ lệ qua đường dẫn từ dưới lên (bottom-up pathway) để phát hiện ROI.

Lơi ích:

* **Phát hiện đối tượng nhỏ**: Bốn đầu ra cải thiện khả năng nhận diện ROI nhỏ, như khối u hoặc tế bào.
* **Tăng cường kết hợp đặc trưng**: Tích hợp hiệu quả thông tin cấp cao và cấp thấp, nâng cao độ chính xác trong các trường hợp ROI đa dạng.
* **Hiệu quả với ảnh y tế**: Phù hợp với các ROI có kích thước không đồng đều, hỗ trợ phát hiện chính xác trong CT hoặc MRI.

## 4.5. Kết quả

Với những cải tiến trên, chúng ta thu được mô hình LSM-YOLO nhẹ hơn với số lượng tham số và pháp tính toán giảm đi đáng kể mà vẫn giữ được hiệu suất cao.

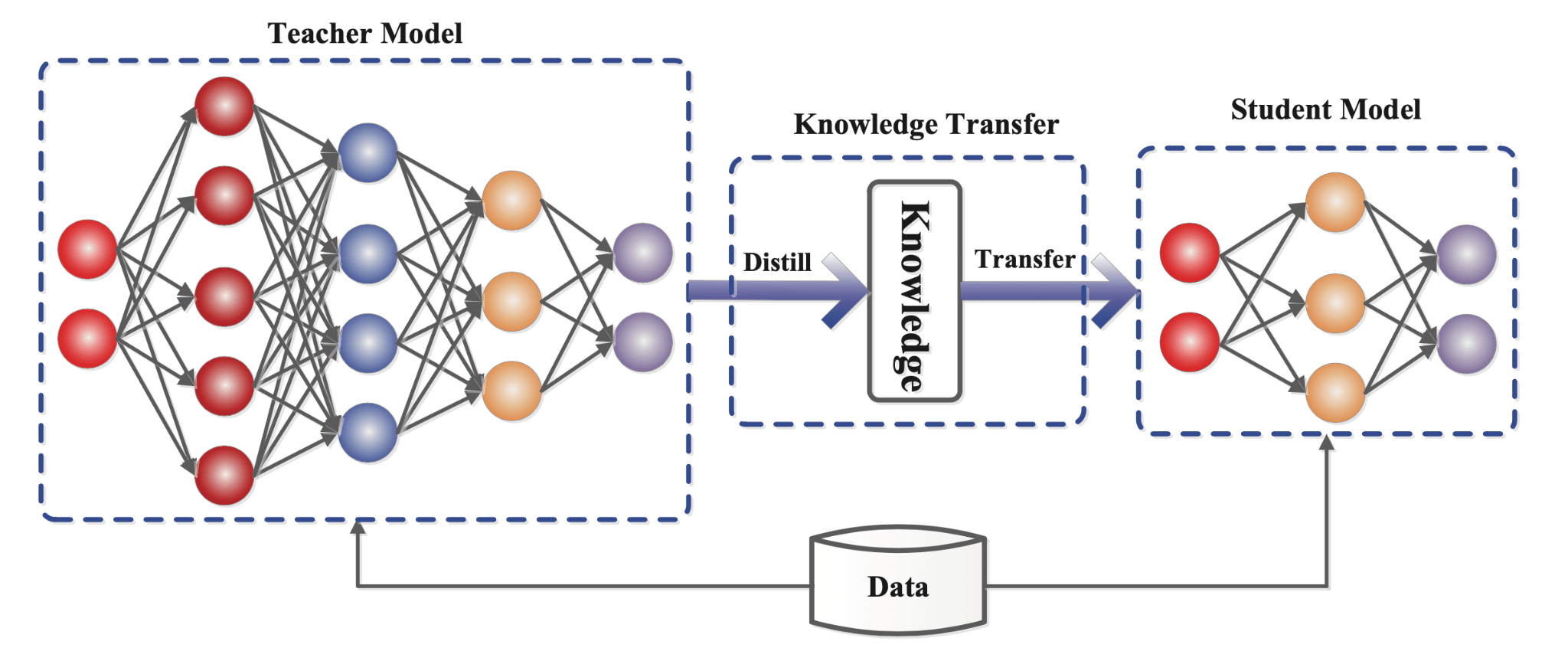
* Phát hiện chính xác ROI nhỏ và chồng lấn nhờ kết hợp đặc trưng đa tỷ lệ hiệu quả (MSFM, PA-FPN).
* Học mối quan hệ không gian giữa ROI và vùng lân cận, phù hợp với khối u xâm lấn (MSFM, RFAConv).
* Giảm nhiễu, tối ưu trên tập dữ liệu y tế nhỏ (LAE).
* Định vị chính xác ROI phức tạp với hàm loss kết hợp (BCE + DFL + SIoU).

# Chương 5. Phương pháp đề xuất

## 5.1. Giới thiệu về kỹ thuật Knowledge Distillation

Trên thực tế, việc triển khai một mô hình học máy trên các thiết bị biên gặp một thách thức lớn đó chính là bộ nhớ và khả năng tính toán. Chúng ta thường giải quyết vấn đề thông qua việc nén mô hình, nhưng lại đồng nghĩa với việc đánh đổi một phần độ chính xác. Việc cân bằng giữa độ chính xác và độ nén lại là một vấn đề khó khăn. Một số phương pháp nén mô hình hay được sử dụng như cắt tỉa, lượng tử hóa đều phải chấp nhận giảm độ chính xác trong một vài trường hợp.

Để giải quyết thách thức trên, một phương pháp nén mô hình mới, chuyển giao tri thức từ một mô hình lớn sang một mô hình nhỏ hơn mà không có bất kỳ tổn thất đáng kể nào. Quá trình một mô hình nhỏ học từ một mô hinhflowns với độ chính xác cao được gọi là “Knowledge Distillation” được giới thiệu bởi Hinton và các cộng sự [[9]](#ref9).



Hình . Knowledge Distillation

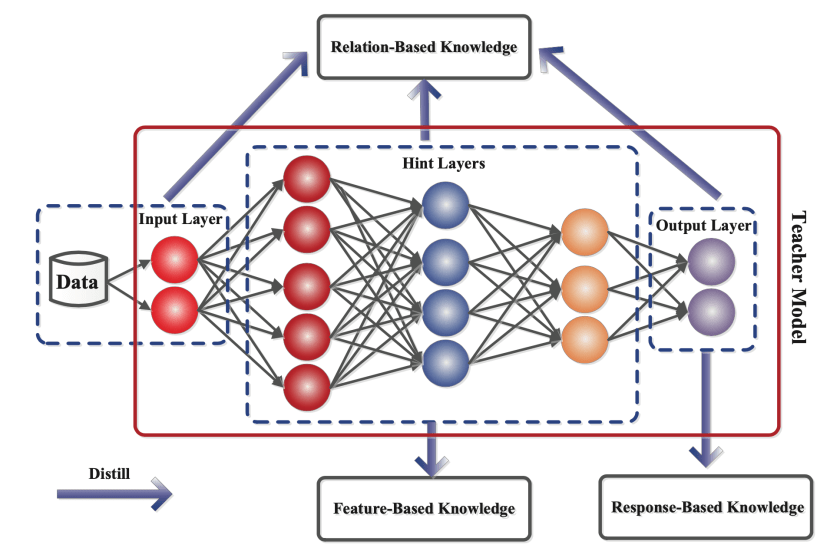
**Knowledge Distillation,** hay Chưng cất tri thức, là quá trinhg chuyển giao kiến thức từ một mô hình lớn, phức tạp hơn (**teacher**) sang một mô hình nhỏ, đơn giản hơn (**student**). Mục tiêu là giúp mô hình nhỏ đạt được hiệu suất gần tương đương với mô hình lớn nhưng với chi phí tính toán nhỏ hơn nhiều, thích hợp cho triển khai trên các thiết bị biên như thiết bị di động, hệ thống nhúng.

### 5.1.1. Các hình thức Knowledge Distillation

#### 5.1.1.1. Knowledge

Như được biết trong các mạng nơ-ron nhân tạo, kiến thức ngoài các trọng số (weigths) và độ lệch (bias) mà còn là:

* Logits: đầu ra trước khi softmax
* Activation: giá trị kích hoạt của các lớp trung gian
* Quan hệ giữa các mẫu
* Tham số của mô hình teacher

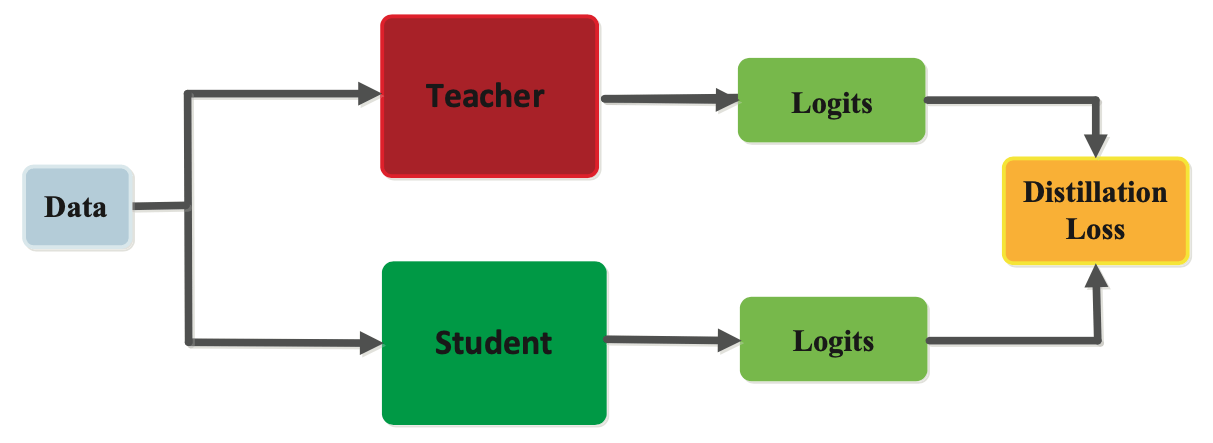


Hình . Các loại tri thức khác nhau trong mạng nơ-ron

Do đó, kiến thức được chia làm 3 loại chính [[10]](#ref10):

1. Response-Based Knowledge:

Đây là kiến thức thường được nhắc đến trong mạng nơ-ron, chính là lớp đầu ra của mô hình teacher. Mục tiêu là giúp mô hình student bắt chước dự đoán của mô hình teacher. Điều này được thực hiện thông qua hàm mất mát chưng cất (Distillation Loss), nó sẽ giúp ta nắm bắt được sự khác nhau giữa logits đầu ra của mô hinhd student và mô hình teacher. Việc tối thiểu hóa hàm mất mát này trong quá trình huấn luyện sẽ giúp các dự đoán của mô hình student giống với mô hình teacher.



Hình . Response-based knowledge distillation

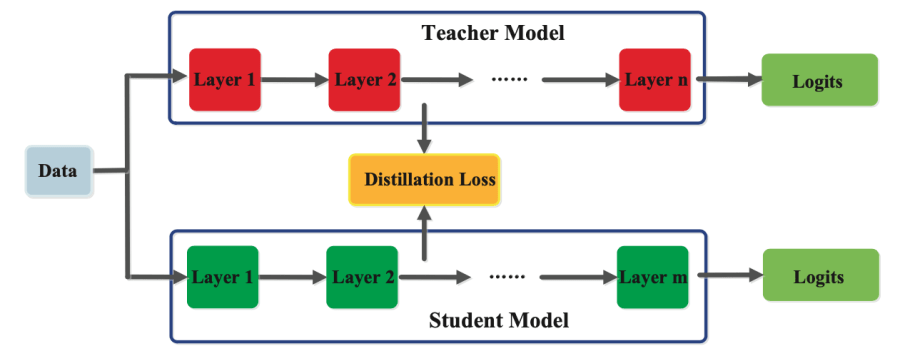
**Distillation Loss**:

Trong đó:

* chính là hàm mất mát giữa nhãn thức tế và đầu ra của mô hình trò (giúp học từ dữ liệu thực).
* là độ sai khác nhau giữa xác suất làm mềm (soft targets) của teacher và student .
* là các hệ số giúp cần bằng giữa việc học từ thực tế và học từ teacher.

1. Feature-Based Knowledge:

Feature-based Knowledge Distillation (KD) là phương pháp chuyển giao kiến thức từ mô hình teacher sang student bằng cách sử dụng đặc trưng trung gian (feature maps) từ các tầng ẩn của teacher, thay vì chỉ dựa vào đầu ra cuối như Response-Based KD.



Hình , Feature-based knowledge distillation

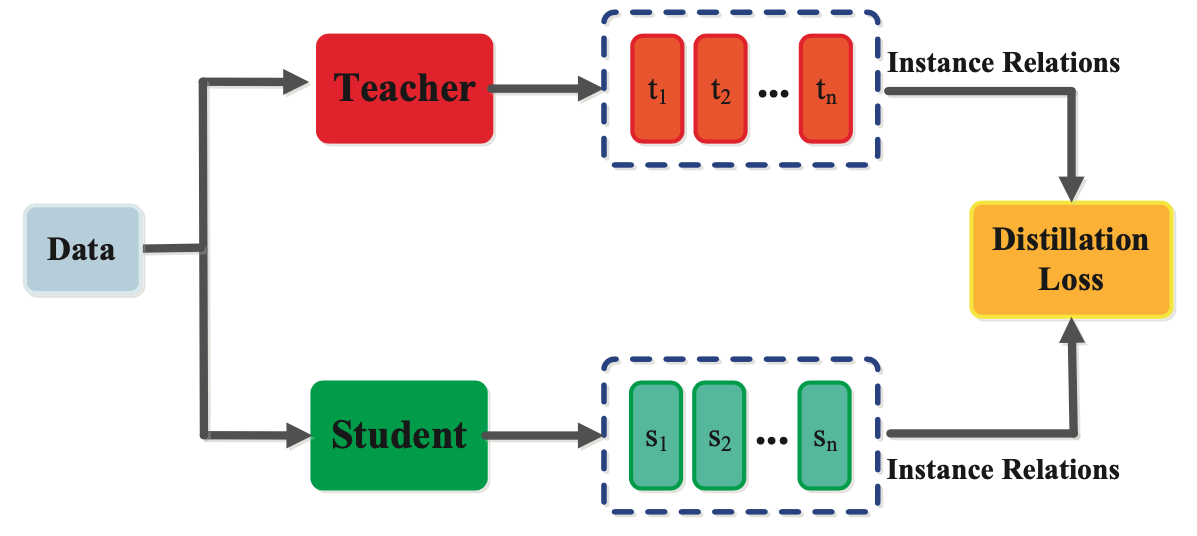
* **Ý tưởng chính:** Student học cách tái tạo các đặc trưng (patterns) mà teacher rút trích từ dữ liệu, giúp nắm bắt thông tin không gian và ngữ nghĩa chi tiết hơn. Ví dụ, trong LSM-YOLO, feature maps có thể chứa thông tin về vị trí và hình dạng của ROI trong ảnh y tế.
* **Cơ chế:** 
  + Teacher cung cấp các feature maps từ các tầng trung gian (thường là backbone hoặc neck trong YOLO).
  + Student được huân luyện để khớp với các feature maps này thông qua hàm mất mát, thường là hoặc :

trong đó là các feature maps của teacher và student.

* **Lợi ích**: Giúp student học được các đặc trưng phức tạp, cải thiện khả năng phát hiện các đối tượng nhỏ hoặc chồng lấn (điều này rất quan trọng trong ảnh y tế). Ngoài ra, giúp cho student không chỉ học từ đầu ra mà còn học được từ những tầng ẩn của teacher.

1. Relation-Based Knowledge:

Relation-based Knowledge Distillation (KD) là phương pháp chuyển giao kiến thức từ mô hình teacher sang student bằng cách tập trung vào mối quan hệ giữa các đặc trưng hoặc mẫu dữ liệu trong teacher, thay vì chỉ sử dụng đầu ra hoặc đặc trưng riêng lẻ.



Hình . Relation-based knowledge distillation

* Ý tưởng chính: Student học cách mô phỏng các tương quan hoặc cấu trúc quan hệ mà teacher rút trích, ví dụ: mối quan hệ giữa các feature maps hoặc giữa các mẫu dữ liệu (như sự tương đồng giữa các ROI trong ảnh y tế).
* Cơ chế:
  + Teacher cung cấp thông tin về quan hệ, như ma trận tương quan giữa các feature maps (FSP matrix) hoặc độ tương đồng giữa các mẫu dữ liệu.
  + Student được huấn luyện để khớp các quan hệ này, sử dụng hàm mất mát đo độ tương đồng

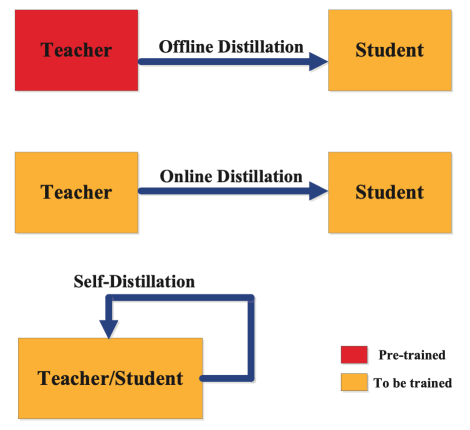
Trong đó, là các hàm tính quan hệ từ đặc trưng của teacher và student, thường là hoặc .

* Lợi ích: giúp student nắm bắt cấu trúc phức tạp và mối quan hệ giữa các đối tượng, phù hợp với cá tác vụ cần hiểu bối cảnh như phân loại videos, nhận diện hành động.

#### 5.1.1.2. Distillation

Có ba cách để thực hiện huấn kuyeehb mô hình student và mô hình teacher. Chúng được phân loại dựa vào cách mà mô hình teacher được sử dụng tại cùng thời điểm với student hay không.

* Offline Distillation: Mô hình teacher được huấn luyện trước, cố định, chuyển kiến thức sang student.
* Online Distillation: Teacher và student huấn luyện đồng thời, học lẫn nhau, nhưng tốn tài nguyên, ít phù hợp với dữ liệu y tế nhỏ.
* Self-Distillation: Mô hình tự học từ chính nó, dùng tầng sâu hướng dẫn tầng nông nhưng hiệu quả hạn chế.

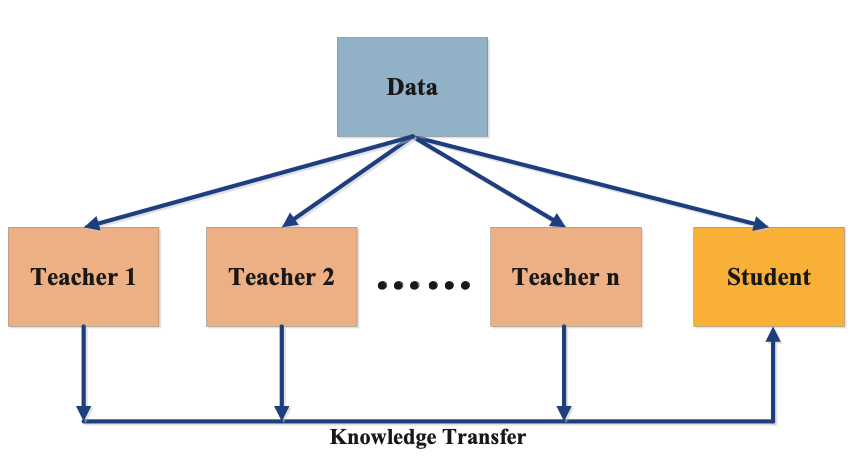


Hình . Types of knowledge distillation training schemes

### 5.1.2. Thuật toán Knowledge Distillation

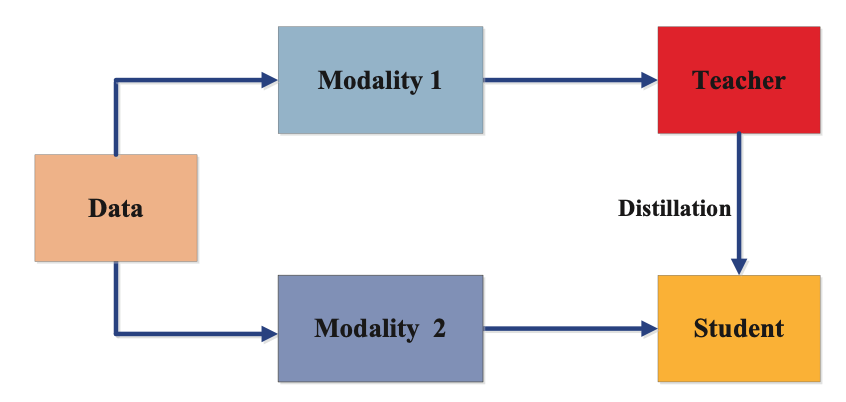
Trong phần này, chúng ta sẽ cùng tìm hiểu vầ các phướng pháp và kỹ thuật được sử dụng để có thể truyền tải các kiến thức từ mô hình teacher sang mô hinh student.

* **Adversarial Distillation** (Chưng cất Đối nghịch): lấy cảm hứng từ Mạng Sinh Đối Nghịch. Sử dụng một bộ phân biệt (discriminator) để đánh giá xem đầu ra hoặc đặc trưng của mô hình sinh viên có "thật" (giống của giáo viên) hay không. Mô hình sinh viên (đóng vai trò như bộ sinh - generator) cố gắng tạo ra các biểu diễn "đánh lừa" được bộ phân biệt. Mục đích giúp cả giáo viên và học sinh hiểu rõ hơn về phân phối dữ liệu thực sự.
* **Multi-Teacher Distillation**: Sử dụng nhiều mô hình teacher khác nhau để hướng dẫn một mô hình student. Các teacher này có thể có kiến trúc mạng khác nhau, được huấn luyện trên các tập data hoặc tác vụ khác nhau.



Hình . Multi-teacher distillation

* **Cross-Modal Distillation**: truyền tải kiến thức giữa các phương thức dữ liệu khác nhau



Hình . Cross-modal distillation

* **Graph-Based Distillation**: sử dụng caaus trúc đồ thị đè biểu diễn và truyền tải tri thức. Mục đích truyển tải dữ liệu có dạng cấu trúc.
* **Attention-Based Distillation**: sử dụng cơ chế chú ý để xác định và truyền tải những phần quan trọng nhất của đặc trưng hoặc thông tin từ teacher sang học sinh
* **Data-Free Distillation**: thức hiện chung cất tri thức mà không cần truyền vào bộ dữ liệu gốc của mô hình teacher. Điều này đặc biệt trong các tính hướng liên quan đến quyền riêng tư, bảo mật của dữ liệu.

A diagram of a diagram

Description automatically generated

Hình . Data-Free Distillation

* **Quantized Distillation**: kết hợp chung cất tri thức với lượng tử háo mạng Nơ-ron (giảm dộ chính xác của bias và weigth). Như vậy, có thể tạo ra được mô hình không những nhỏ về tham số mà còn nhỏ về khối lượng tính toán.
* **Lifelong Distillation**: mô hình không chỉ học tác vụ mới mà không quên tác vụ cũ
* **NAS-Based Distillation**: kết hợp chưng cất và Tìm kiếm kiến trúc mạng nơ-ron phù hợp cho mô hình student.

## 5.2. Áp dụng kỹ thuật Knowledge Distillation với LSM-YOLO

Về việc áp dụng kỹ thuật Knowledge Distillation cho LSM-YOLO. Ta xét ngữ cảnh huấn luyện. Chúng ta tập trung vào bộ dữ liệu Br35H đã được tiền xử lý, tạp trung vào phát hiện vùng quan tâm ROI trong ảnh y tế. Ảnh y tế với đạc điểm nhỏ, có sự chồng lấn và đặc biệt là sự tương quan giữa các vùng mô. Đây là đặc điểm nổi bật và đặt trưng của loại dữ liệu này. Xét trên ngữ cảnh đó, chúng em chưng cất tri thức để cải thiện LSM-YOLO

### 5.2.1. Mô hình teacher

Trong việc chọn mô hình teacher cho LSM-YOLO, tham khảo qua nhiều mô hình khác nhau, chúng em quyết định chọn hai model YOLOv8s và YOLOv9c. Cả hai model trên đều có nhưng đặc điểm phù hợp với LSM-YOLO và phù hợp với bộ dữ liệu y tế.

* Hiệu suất và chất lượng Feature Maps
  + YOLOv8-S là phiên bản nhỏ của YOLOv8 nổi bật với hiệu suất cao nhờ các cải tiến như C2f Modules và SPPF (Spatial Pyramid Pooling – Fast). Nó tạo ra các feature maps với chất lượng cao ở các tần trung gian.
  + YOLOv9-C sử dụng Programmable Gradient Information (PGI) và GELAN, tạo feature maps đa tỷ lệ phong phú, lý tưởng cho phát hiện các đối tượng phức tạp
  + VỚi các feature maps chất lượng giúp cho LSM-YOLO học tốt hơn về vị trí và kích thước vùng ROI trong ảnh y tế.
* Độ tương thích với ảnh y tế
  + YOLOv8-S mạnh trong phân loại, phù hợp để nhận diện vùng ROI trong ảnh y tế
  + YOLOv9-C vượt trội trong phát hiện đối tượng đa tỷ lệ, rất hiệu quả cho các ROI nhỏ hoặc chồng lấn trong ảnh y tế như khối u não
  + Đặc biệt cả hai đều đã đượcc huấn luyện trên bộ dữ liệu COCO, có khả năng tổng quát hóa cao, giúp LSM hoạt động tốt trên các bộ dữ liệu y tế nhỏ và nhiều biến động
* Độ tương thíc với LSM-YOLO:
  + LSM-YOLO là mô hình nhẹ (2.87M tham số, 12.4 GFLOPs), cần mô hình teacher mạnh nhưng không quá phức tạp. YOLOv8-S (nhẹ hơn) và YOLOv9-C (cân bằng) phù hợp, tránh chênh lệch kiến trúc lớn so với các mô hình nặng hơn (như YOLOv9-E.

### 5.2.2. Knowledge

Việc lựa chọn kiến thức dựa trên đặc trưng (feature-based knowledge) để chưng cất cho mô hình LSM-YOLO là một quyết định chiến lược nhằm tối ưu hóa hiệu suất của mô hình sinh viên nhỏ gọn này. Thay vì chỉ bắt chước đầu ra cuối cùng, phương pháp này cho phép LSM-YOLO học hỏi sâu hơn về cách mô hình giáo viên lớn hơn xử lý và biểu diễn dữ liệu ở các lớp trung gian, cung cấp những "gợi ý" phong phú và chi tiết. Điều này đặc biệt quan trọng vì LSM-YOLO, với năng lực biểu diễn hạn chế hơn, có thể gặp khó khăn khi tự mình khám phá các đặc trưng phức tạp. Bằng cách chưng cất đặc trưng từ các tầng then chốt của giáo viên, chúng ta giúp LSM-YOLO thu hẹp khoảng cách năng lực, học cách tạo ra các biểu diễn mạnh mẽ ở từng giai đoạn của kiến trúc phân cấp YOLO. Quá trình này không chỉ hướng dẫn việc tìm kiếm trong không gian tham số một cách hiệu quả hơn mà còn cải thiện khả năng tổng quát hóa của LSM-YOLO, vì nó học được "quy trình tư duy" của giáo viên chứ không chỉ là kết quả cuối cùng. Hơn nữa, kiến thức đặc trưng giúp khai thác hiệu quả kiến thức ngầm mà giáo viên đã tích lũy, cho phép LSM-YOLO "đứng trên vai người khổng lồ" để đạt được hiệu suất cao hơn. Cuối cùng, việc sử dụng các lớp adapter còn tăng cường tính linh hoạt, cho phép chưng cất hiệu quả ngay cả khi có sự khác biệt về số kênh đặc trưng giữa giáo viên và sinh viên.

### 5.2.3. Kiến trúc teacher-student

Vì hạn chế về mặt phần cứng cùng như đã có các model teacher đã được fine tune với kết quả khác tốt trên bộ dữ liệu Br35H nên việc lựa chọn kiểu kiến trúc “Single Teacher-Single Student” là một giải pháp hợp lý trong ngữ cảnh này.

# Chương 6. Kết quả thực nghiệm

## 6.1. Các thông số cho quá trình chưng cất tri thức với LSM-YOLO

Với các kỹ thuật và model teacher được đề cập chi tiết trong [mục 5.2](#_5.2._Áp_dụng), chúng em tiến hành chưng cất tri thức. Môi trường chưng cất được lựa chọn là Kaggle. Với 2 model teacher là YOLOv8-S và YOLO-v9C. Các thông số, siêu tham số được đề cập trong Bảng 2.

|  |  |
| --- | --- |
| Tham số/ Siêu tham số | Giá trị |
| lambda\_kd | 3 |
| [(Main loss)](#main_loss) | 7.5 |
| [(Main loss)](#main_loss) | 0.5 |
| [(Main loss)](#main_loss) | 1.5 |
| Epochs | 300 |
| Batch\_size | 16 |
| Img\_size | 640x640 |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Beta1 (AdamW) | 0.9 |
| Beta1 (AdamW) | 0.999 |
| Weight Decay (Nhóm g0 - Main Weights) | 1e-5 |
| Weight Decay (Nhóm g1 - BN Weights) | 0.0 |
| Weight Decay (Nhóm g2 - Biases) | 0.0 |
| Weight Decay (Nhóm adapters) | 1e-5 |

Bảng 2. Các thông số huấn luyện

## 6.2. Kết quả thức nghiệm.

Kết quả thực nghiệm được trình bày trong Bảng 3 cung cấp một cái nhìn sâu sắc và đầy thuyết phục về hiệu quả của các mô hình phát hiện đối tượng, đặc biệt là sự thành công vượt trội của phương pháp chưng cất tri thức (Knowledge Distillation - KD) được áp dụng để nén các mô hình YOLOv8s và YOLOv9c thành các phiên bản Lightweight Shunt Matching (LSM) nhỏ gọn hơn.

Điểm nổi bật nhất chính là khả năng của kỹ thuật KD trong việc nén các mô hình nhỏ cực kỳ hiệu quả về mặt tài nguyên. Cụ thể, các mô hình "LSM with YOLOv8s" và "LSM with YOLOv9C" đều chỉ sở hữu 2.87 triệu tham số (Params) và yêu cầu 12.4 GFLOPs, một sự cắt giảm đáng kinh ngạc so với các mô hình "teacher" gốc (YOLOv8s với 11.13M Params, 28.4 GFLOPs và YOLOv9c với 25.32M Params, 102.3 GFLOPs). Việc giảm kích thước mô hình xuống chỉ còn khoảng 1/4 đến 1/9 và giảm GFLOPs xuống khoảng 1/2 đến 1/8 so với các mô hình gốc là một cải tiến kỹ thuật đáng chú ý, cho thấy tiềm năng to lớn trong việc triển khai các mô hình AI tiên tiến trên các thiết bị có tài nguyên hạn chế.

Điều đáng nói hơn nữa là dù có sự thu hẹp đáng kể về kích thước và độ phức tạp tính toán, các mô hình LSM được huấn luyện với KD vẫn duy trì được hiệu suất rất đáng nể. "LSM with YOLOv9C" đạt mAP50-95 là 64.40% và mAP-50 là 88.70%. So với "teacher" của nó là YOLOv9c (68.18% mAP50-95, 92.57% mAP-50), sự sụt giảm hiệu suất là chấp nhận được, đặc biệt khi xét đến lợi ích khổng lồ về mặt hiệu quả tính toán. Tương tự, "LSM with YOLOv8s" cũng cho thấy khả năng học hỏi tốt từ YOLOv8s, đạt 62.10% mAP50-95.

Vai trò then chốt của phương pháp chưng cất tri thức được thể hiện rõ ràng nhất khi so sánh với "LSM NO KD". Mô hình LSM được huấn luyện mà không có sự hướng dẫn từ "teacher" (NO KD) chỉ đạt được mAP50-95 là 46.99% và mAP-50 là 72.16%, thấp hơn rất nhiều so với các phiên bản có KD. Sự chênh lệch hiệu suất này (gần 15-18% ở mAP50-95) là một minh chứng cho tính hiệu quả và vai trò không thể thiếu của kỹ thuật chưng cất tri thức trong việc chuyển giao kiến thức từ một mô hình lớn, phức tạp sang một mô hình nhỏ hơn, nhẹ hơn mà vẫn giữ được phần lớn hiệu suất.

Một điểm thú vị khác là "LSM with YOLOv9C" cho kết quả nhỉnh hơn "LSM with YOLOv8s", mặc dù cả hai mô hình "student" này có cùng kiến trúc và kích thước. Điều này có thể gợi ý rằng việc chưng cất từ một "teacher" mạnh mẽ hơn (YOLOv9c có hiệu suất cao hơn YOLOv8s) có thể dẫn đến một mô hình "student" tốt hơn, ngay cả khi "student" đó rất nhỏ gọn.

Tóm lại, các kết quả thực nghiệm đã khẳng định mạnh mẽ sự thành công của phương pháp chưng cất tri thức mà nhóm đã áp dụng. Kỹ thuật này không chỉ giúp giảm thiểu đáng kể gánh nặng tính toán và yêu cầu bộ nhớ của các mô hình YOLO tiên tiến, mà còn đảm bảo rằng các mô hình nén (LSM) vẫn giữ được độ chính xác ở mức cao. Mặc dù có sự đánh đổi nhất định về chỉ số Recall (R), đặc biệt ở các mô hình LSM, lợi ích tổng thể về hiệu quả và khả năng triển khai là không thể phủ nhận.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Params  (M) | GFLOPs | mAP50-95 | mAP-50 | P | R |
| YOLOv8s | 11.13 | 28.4 | 70.07% | 93.83% | 94.79% | 88.03% |
| YOLOv9c | 25.32 | 102.3 | 68.18% | 92.57% | 92.45% | 86.56% |
| LSM with YOLOv8s | 2.87 | 12.4 | 62.10% | 84.50% | 93.90% | 72.00% |
| LSM with YOLOv9C | 2.87 | 12.4 | 64.40% | 88.70% | 92.80% | 76.90% |
| LSM NO KD | 2.87 | 12.4 | 46.99% | 72.16% | 89.47% | 52.69% |

Bảng 3. Kết quả thực nghiệm

# Tài liệu tham khảo

[[1] Md Shakhawat Hossain, Galib Muhammad Shahriar, M. M. Mahbubul Syeed, Mohammad Faisal Uddin, Mahady Hasan, Shingla Shivam & Suresh Advani, Region of interest (ROI) selection using vision transformer for automatic analysis using whole slide images, 2023](https://www.nature.com/articles/s41598-023-38109-6)

[[2] Xiaoyan Jiang, Zuojin Hu, Shuihua Wang, Yudong Zhang, Deep Learning for Medical Image-Based Cancer Diagnosis, 2023](https://pmc.ncbi.nlm.nih.gov/articles/PMC10377683/)

[[3] Yi-Ching Cheng, Yi-Chieh Hung, Guan-Hua Huang, Tai-Been Chen, Nan-Han Lu, Kuo-Ying Liu, Kuo-Hsuan Lin, Deep Learning-Based Object Detection Strategies for Disease Detection and Localization in Chest X-Ray Images, 2024](https://pmc.ncbi.nlm.nih.gov/articles/PMC11640298/)

[[4] Br35H:: Brain Tumor Detection 2020](https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection)

[[5] LSM-YOLO: A Compact and Effective ROI Detector for Medical Detection](https://arxiv.org/abs/2408.14087)

[[6] Gevorgyan, Z.: Siou loss: More powerful learning for bounding box regression. arXiv preprint arXiv:2205.12740 (2022)](https://arxiv.org/abs/2205.12740)

[[7] Zhang, X., Liu, C., Yang, D., Song, T., Ye, Y., Li, K., Song, Y.: Rfaconv: In novating spatital attention and standard convolutional operation. arXiv preprint arXiv:2304.03198 (2023)](https://arxiv.org/abs/2304.03198)

[[8] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8759–8768 (2018)](https://ieeexplore.ieee.org/document/8579011)

[[9] Distilling the Knowledge in a Neural Network](https://arxiv.org/abs/1503.02531)

[[10] Knowledge Distillation: A Survey](https://arxiv.org/pdf/2006.05525)