1. **Giới thiệu đề tài** 
   1. **Lý do chọn đề tài**

Phát hiện đối tượng là bài toán phổ biến của cộng đồng nghiên cứu trí tuệ nhân tạo nói chung và lĩnh vực thị giác máy tính nói riêng. Nhờ vào sự phát triển của của phương pháp học sâu, bài toán phát hiện đối tượng ngày càng được cải tiến và áp dụng trong nhiều lĩnh vực cụ thể, đóng góp vai trò quan trọng trong lĩnh vực đó. Đầu vào của bài toán là hình ảnh/video hay hình ảnh truyền tải từ các camera real-time, đầu ra sẽ là vị trí của các đối tượng cần được xác định. Bên cạnh đó, với sự phát triển của máy bay không người lái thì việc ứng dụng thiết bị này để giải quyết các vấn đề liên quan đến giao thông trở nên vô cùng thu hút, do đó bài toán phát hiện phương tiện giao thông không ảnh nhận được rất nhiều sự quan tâm của cộng đồng khoa học thế giới.

Bên cạnh đó, kinh tế ngày càng phát triển nhu cầu di chuyển và vận tải ngày càng tăng. Số lượng phương tiện giao thông vì thế cũng tăng theo. Lượng xe lớn tạo nhiều áp lực cho các cơ quan quản lý, giám sát giao thông. Các vấn đề như tắc đường, tai nạn giao thông, thống kê phương tiện, quy hoạch các tuyến đường và bãi đỗ xe cũng đang hiển nhiên rất cần những giải pháp kịp thời và mang tính ổn định cao. Một lượng lớn các thiết bị camera, cảm biến, radar được lắp đặt để giám sát xe cộ và thu thập thông tin giao thông, giúp các cơ quan nắm tình hình lưu lượng giao thông, mật độ phương tiện và tình trạng đỗ xe. Tuy nhiên, những phương pháp này không cung cấp đủ cái nhìn tổng quan về tình hình giao thông, mà đây lại là thông tin quan trọng để đưa ra các giải pháp phát triển và giải quyết tình trạng hiện tại.

Gần đây, hình ảnh chụp từ UAVs được ưu tiên hơn do khả năng bao quát cả khu vực và độ phân giải không gian cao hơn từ 0.1 đến 0.5m [1] và dễ thu thập hơn. Với độ phân giải cao, các phương tiện giao thông dễ dàng được phát hiện, kể cả đối tượng nhỏ như ôtô, xe máy. Việc phát hiện phương tiện giao thông với ảnh chụp từ UAVs tồn tại nhiều thách thức do sự xuất hiện của nhiều đối tượng khác như dây điện, cây xanh, tòa nhà, máy điều hòa không khí, bảng hiệu, thùng rác có thể gây nhiễu và đưa ra các cảnh báo sai. Đặc biệt là điều kiện ánh sáng, khi mà việc xuất hiện bóng của các tòa nhà, phương tiện giao thông, cây cối gây ra rất nhiều khó khăn trong công tác nhận dạng và phát hiện đối tượng phương tiện

Bài toán tồn tại nhiều thách thức do sự đa dạng về góc nhìn, hình dáng , kích thước của đối tượng.Đặc biệt, trong không gian ảnh , các hình ảnh được thu từ trên không nên xuất hiện nhiều thách thức mới như hình ảnh có thể được chụp từ các độ cao khác nhau : (1)ảnh độ phân giải cao , các đối tượng chiếm tỉ lệ đa dạng , phân bố thưa hoặc tập trung dày đặc; (2)Sự mất cân bằng giữa các đối tượng trong các lớp và giữa foreground và background do độ cao khi chụp. (3)Tính di động của flycam ,cũng như camera gắn kèm ,xuất hiện nhiều góc nhìn làm cho cùng một đối tượng ( phía trước, phía sau, bên phải ,… hay hướng các đối tượng tùy ý ) lại tạo nên đa dạng các thể hiện .

(4) Điều kiện thời tiết , ánh sáng ( trời nắng hay mưa , có sương mù , ban ngày , ban đêm ,…) ảnh hướng đến khả năng hiển thị đối tượng trong ảnh .

Để giải quyết tốt bài toán phát hiện phương tiện giao thông, cần có một mô hình phân lớp tốt. Mà, để có được một mô hình phân lớp tốt, cần biết cách trích chọn đặc trưng đầu vào sao cho mô hình cuối cùng có thể học được nhiều đặc trưng ý nghĩa nhất. Mặt khác, bộ dữ liệu được đưa vào để huấn luyện cũng phải đa dạng nhưng đồng thời cũng phải phù hợp với ngữ nghĩa, hoàn cảnh giao thông Việt Nam mới có thể giải quyết được vấn đề.

Chính vì những thách thức trên của bài toán nhận diện phương tiện , nhóm chúng em đã quyết định tiến hành tìm hiểu về nhận diện phương tiện, đồng thời tìm hiểu một vài phương pháp thực nghiệm như Faster R-CNN, Double Head ,… để từ đó tìm ra phương pháp khắc phục những thách thức của bài toán phát hiện phương tiện tham gia giao thông.

Trong quá trình thực hiện đề tài , nếu nhóm chúng em có sai sót thì mong nhận được sự thông cảm và góp ý từ thầy để hoàn thiện đồ án tốt hơn. Xin cảm ơn thầy đã dành thời gian để nhóm em được trình bày về đồ án của nhóm.

**1.2. Mục tiêu hướng đến**

- Khảo sát về bài toán Object detection , cụ thể thông qua bài toán phát hiện đối tượng trong không ảnh

- Đánh giá và đề xuất các phương pháp học sâu (Deep learning ) được sử dụng trong đồ án. Bao gồm Faster R-CNN ( cùng với Feature Pyramid Network) , Double Head , GroIE, Guided Anchoring, Side-aware Boundary Localization (SABL)

- Đồng thời đưa ra phương pháp đề xuất là kết hợp GDSNet (GRoIE + Double Head + SABL) , sau đó dựa vào kết quả thu được để đánh giá hiệu quả khi kết hợp.

**2. Phương pháp thực nghiệm**

**2.1 : thu thập dữ liệu**

**2.2**

**2.3 : kiến trúc CNN dùng để trích xuất đặc trưng**

**2.3.1 : resnet( resnet50 và resnet 101)**

ResNet, là viết tắt của Residual Networks, là một kiến trúc CNN được đề xuất bởi He và cộng sự vào năm 2015[11]. Đây là kiến trúc đã huấn luyện ra mô hình đạt giải nhất cuộc thi ImageNet. Khác với các kiến trúc CNN truyền thống trước đây, ResNet sử dụng các khối residual để tránh tình trạng Vanishing Gradient (các trọng số không còn được cập nhật ở các lớp gần cuối).

Graphical user interface, diagram, application

Description automatically generated

**Hình 1.6 Khối thông thường và khối residual [19]**

Có thể quan sát ở Hình 1.6, một khối thông thường ta sẽ thấy nhiều ở các kiến trúc CNN trước đây, đó là từ một đầu vào x, sau khi đi qua các lớp tích chập và các hàm kích hoạt sẽ cho ra một đầu ra f(x), đầu ra này tiếp tục đi qua một hàm kích hoạt nữa. Tuy nhiên đối với khối residual bên phải, đầu ra của nó sẽ được cộng thêm với đầu vào x, rồi sau đó mới đi qua hàm kích hoạt. Cơ chế này được gọi là skip connection (kết nối tắt).

Các lợi ích của skip connection như sau:

* Giải quyết vấn đề vanishing gradient bằng cách sử dụng các kết nối tắt.
* Đảm bảo các lớp sâu hơn sẽ hoạt động ít nhất là tốt hơn các lớp gần ảnh và không tệ hơn.
* Trong đề tài này, tôi sẽ huấn luyện Faster R – CNN với ResNet-50 và ResNet-101, tức là kiến trúc ResNet nhưng bao gồm 50 lớp tích chập:

Diagram

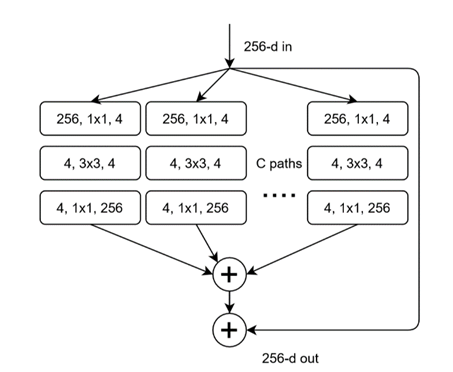
Description automatically generated

**Hình 1.7 Kiến trúc ResNet50 [19]**

Quan sát ở Hình 1.7 có thể thấy rằng ResNet-50 bao gồm các khối residual kết hợp lại, trong mỗi khối residual bao gồm 3 lớp tích chập: 2 lớp có kernel 1×1 và 1 lớp 3×3. Trong đó 2 lớp đầu của mỗi khối có số kênh như nhau, còn lớp cuối thì số kênh sẽ gấp bốn lần lên. Sau mỗi khối sẽ là các lớp max pooling để giảm độ phân giải đặc trưng xuống hai lần. Ta có tổng cộng 1 lớp 64 kênh ở đầu kiến trúc, 3 khối đầu ra 256, 4 khối đầu ra 512, 6 khối đầu ra 1024 và 3 khối cuối đầu ra 2048, các đặc trưng sau khi thu được ở ResNet sẽ qua các bước xử lý tiếp theo trong Faster R – CNN. Mạng ResNet-101 cũng có kiến trúc tương tự tuy nhiên sẽ có nhiều lớp hơn.

2.3.2 :Resnext

Được đề trong nghiên cứu [12], kiến trúc ResNeXt cũng sử dụng cơ chế kết nối tắt tuy nhiên các tác giả chia thành nhiều nhánh trong khối residual, sau đó các nhánh sẽ được cộng lại. Số nhánh được chia ra sẽ là một siêu tham số gọi là “cardinality”, “cardinality” càng cao sẽ giúp giảm lỗi trên tập kiểm định, điều đó có nghĩa là việc overfitting sẽ được giải quyết. Số chiều của mỗi nhánh được ký hiệu là d, hình ảnh minh họa của Residual block trong ResNeXt được thể hiện ở Hình 1.8.



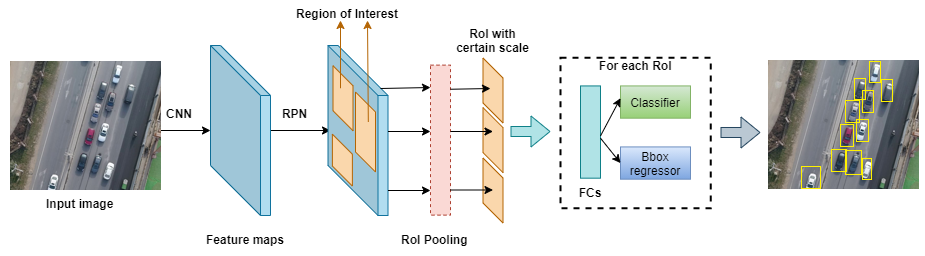
Hình 1.8 : minh họa của Residual block trong ResNeXt

**2.3.2 : Swin transformer**

**2.4 :Các phương pháp thực nghiệm**

**2.4.1 Faster R-CNN**

Faster R – CNN [25] là phiên bản cải tiến của Fast R – CNN [36]. Ren cùng cộng sự đã đề xuất một mạng đề xuất khu vực (Region Proposal Network), thay thế selective search để tạo ra các vùng đề xuất tốt hơn, kiến trúc này sau đó sẽ được huấn luyện cùng với Fast R – CNN. Những cải tiến này đã làm giảm số lượng các vùng đề xuất và tăng tốc độ hoạt động trong thời gian thử nghiệm mô hình lên gần thời gian thực với hiệu suất tốt nhất, tốc độ khoảng 5fps trên một GPU. Faster R – CNN là phương pháp tiền đề cho rất nhiều phương pháp phát hiện đối tượng sau này.



**Hình Framework của Faster R – CNN**

Hình 2.8 mô tả tổng quan Faster R – CNN. Một ảnh đầu vào sẽ đi qua một kiến trúc CNN và có đầu ra là đặc trưng của ảnh (feature map). Đặc trưng này sau đó đi qua RPN để sinh ra các vùng đề xuất có hoặc không có đối tượng, các vùng này sẽ qua lớp RoI Pooling để được resize lại cùng một kích thước cho trước, sau đó sẽ được phân loại và dự đoán tọa độ bởi Fast R – CNN.

Cụ thể, Region Proposal Networks (RPN) sẽ trượt một cửa sổ trượt trên feature map và sẽ sinh ra vùng đề xuất nằm trong sửa sổ đó. Các vùng đề xuất này sau đó sẽ được tham số hóa thành các hộp neo (anchor box).

Để huấn luyện RPN, mỗi anchor box sẽ được gán nhãn là dương tính (positive) hoặc là âm tính (negative) dựa vào độ đo IoU lớn nhất giữa anchor box với tất cả hộp bao ground truth. Gọi là anchor box đang xét, là tập ground truth, ta có công thức xác định sau:

Các trường hợp còn lại sẽ bị bỏ qua.

Sau khi gán nhãn, các anchor box sẽ được đưa vào RPN để huấn luyện. Các vùng đề xuất sau đó sẽ được resize về một kích thước và cùng đi qua hai lớp Fully Connected (FC). Trong đó, lớp FC đầu tiên sẽ được dùng để phân lớp nhị phân (có hoặc không có đối tượng) có đầu ra là 2, lớp FC thứ hai được dùng để hồi quy tọa độ của đối tượng với đầu ra là 4, ứng 4 giá trị trong tọa độ của vùng đề xuất đối tượng. Hàm mất mát dùng để huấn luyện RPN là sự kết hợp của hàm mất mát phân loại nhị phân có hoặc không có đối tượng và hàm mất mát hồi quy tọa độ bounding box, được định nghĩa như sau:

Điều đặc biệt của Faster R – CNN đó là cả 2 mạng RPN và Fast R – CNN sử dụng chung vài lớp tích chập và sử dụng chung toàn bộ đặc trưng ảnh. Trong bài báo gốc, Ren cùng cộng sự đã sử dụng chiến lược huấn luyện luân phiên để thực hiện các thử nghiệm:

Bước 1: Đầu tiên, các lớp dùng chung sẽ sử dụng trọng số huấn luyện trước trên bộ ImageNet, trọng số các lớp riêng của RPN sẽ được khởi tạo ngẫu nhiên.

Bước 2: Huấn luyện RPN, sau khi RPN dự đoán các vùng đề xuất, trọng số của các lớp riêng RPN và các lớp dùng chung sẽ được cập nhật.

Bước 3: Ở bước huấn luyện Fast R – CNN, ban đầu các lớp dùng chung sẽ có trọng số như ở Bước 2, còn trọng số các lớp riêng của Fast R – CNN sẽ được khởi tạo ngẫu nhiên.

Bước 4: Huấn luyện Fast R – CNN, lúc này cả các lớp dùng chung và các lớp riêng của Fast R – CNN sẽ được cập nhật lại. Bộ trọng số cập nhật này tiếp tục được sử dụng RPN và lặp lại quá trình như trên.

**2.4.2. Feature Pyramid Network**

Mạng đặc trưng đa tầng (FPN) [13] là một bộ trích xuất đặc trưng nhận một hình ảnh tỷ lệ đơn có kích thước tùy ý làm đầu vào và đầu ra là các đặc trưng có kích thước tương ứng ở nhiều tầng. FPN được đề xuất nhằm kết nối các đặc trưng ngữ nghĩa yếu với các đặc trưng ngữ nghĩa mạnh, và quá trình này sẽ độc lập với các kiến trúc mạng tích chập xương sống truyền thống. FPN sẽ được kết hợp với kiến trúc ResNet và ResNeXt để huấn luyện Faster R-CNN và Cascade R-CNN nhằm giúp kết quả phát hiện vị trí đạt hiệu suất tốt nhất.

Diagram

Description automatically generated

**Hình Minh họa Feature Pyramid Network [21]**

Được thể hiện ở Hình 1.9, FPN bao gồm hai thành phần: đường từ dưới lên trên (Bottom-Up Pathway) và đường từ trên xuống dưới (Top-Down Pathway). Bên cạnh có cơ chế kết nối bên (lateral connection). Sau đây tôi giải thích rõ các thành phần này.

**Đường từ trên xuống (Bottom-Up Pathway)**: minh họa cho các kiến trúc mạng CNN hiện nay, mà qua đó một ảnh đầu vào đi qua các lớp tích chập sẽ thu được các đặc trưng đầu ra tương ứng với các lớp đó. Các đặc trưng đầu ra này (bên trái) sau đó sẽ được kết hợp với các đặc trưng ở Top-Down Pathway qua một kết nối bên (lateral connection).

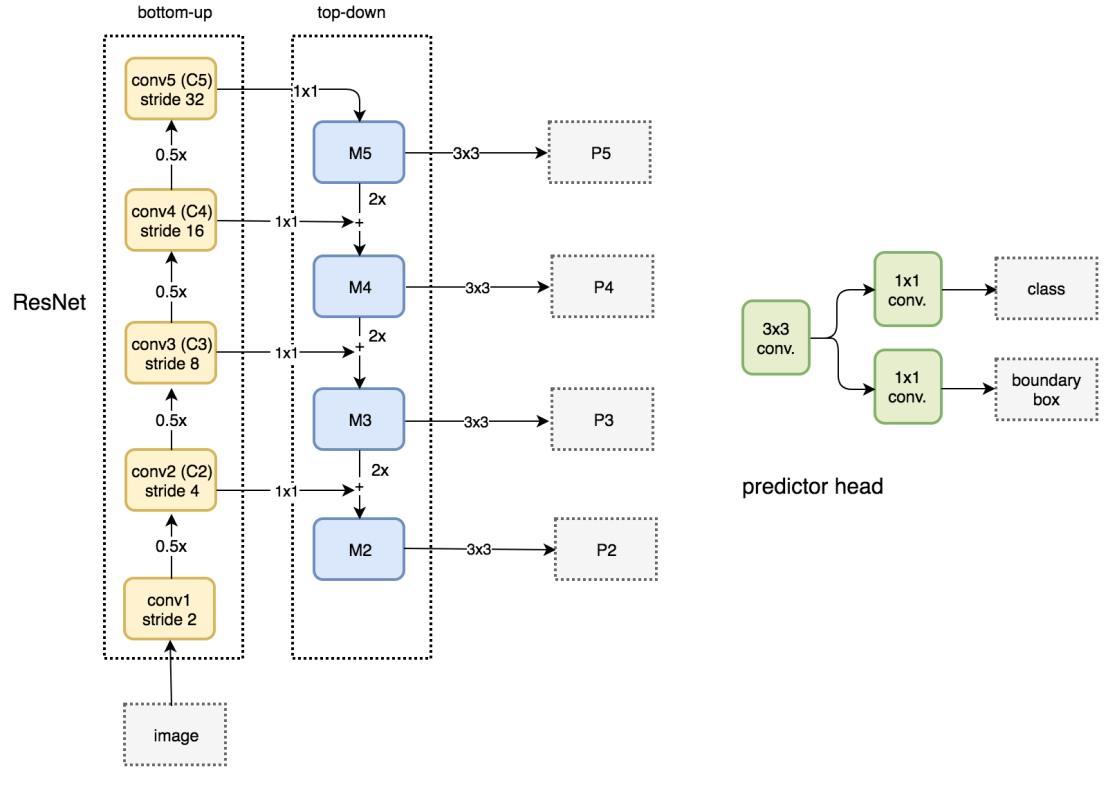
**Đường từ dưới lên và Kết nối bên (Top-Down Pathway)**: được đề xuất với một nỗ lực tăng kích thước của các đặc trưng ngữ nghĩa cao này. Cụ thể, đặc trưng ở đỉnh kim tự tháp bên trái sẽ được tăng độ phân giải lên hai lần (upsampling) để tăng chất lượng lên. Tuy nhiên có thể việc tăng độ phân giải như vậy sau nhiều lớp sẽ làm mất đi thông tin ảnh. Do đó cần một kết nối bên để kết hợp đặc trưng bên trái ứng với tầng của nó bên phải. Trong kết nối bên có các thành phần:

1x1 conv: đây là một lớp tích chập thông thường, để đảm bảo đặc trưng bên trái cùng số chiều (channel) với đặc trưng bên phải cùng tầng với nó.

2x up: phép upsampling tăng độ phân giải đặc trưng lên hai lần.

Sau khi thực hiện xong tích chập đặc trưng bên trái, cũng như tăng độ phân giải đặc trưng bên phải, ta thực hiện phép cộng, và cứ tiếp tục như vậy, các đặc trưng này sau đó còn đi qua một lớp tích chập với kernel nữa để thu được các đặc trưng sau cùng. Các đặc trưng đa cấp đầu ra cuối cùng sẽ được sử dụng để làm đầu vào cho các bộ phát hiện đối tượng.

Sử dụng FPN trong mạng đề xuất vùng trong Faster R-CNN và Cascade R-CNN (RPN): Trong bài báo gốc, RPN chính là một lớp tích chập sử dụng kernel 3×3 và theo sau nó là hai lớp tích chập kernel 1×1 dùng để lần lượt phân loại có hoặc không phải đối tượng và hồi quy tọa độ, và RPN được đặt ở trên đầu kiến trúc CNN tỷ lệ đơn. Ở đây, đầu kiến trúc CNN tỷ lệ đơn này sẽ được thay thế bởi FPN. FPN sẽ cho ra nhiều đặc trưng vì nó có nhiều tầng, các đặc trưng P2,P3,P4,P5 này sẽ lần lượt đi qua RPN (Hình 1.10). Bởi vì đầu ra của FPN là nhiều tầng, do đó nó sẽ không cần thiết phải nó nhiều hộp mỏ neo ở nhiều tỷ lệ các nhau trên một tầng nhất định, mà một tỷ lệ đơn sẽ được gán cho một tầng.



**Hình minh họa việc áp dụng FPN với RPN [21]**

Sử dụng FPN trong mạng detection: Trước đây, đặc trưng sau khi qua RPN sẽ nhận được các vùng đề xuất, các vùng đề xuất này sau đó sẽ đi qua RoI pooling để đưa về cùng một kích thước. Khi áp dụng FPN, đầu ra của nó sẽ là các đặc trưng nhiều tầng, sau đó áp dụng RPN để tạo ra các RoI (các vùng đề xuất). Dựa vào kích thước của mỗi vùng, ta sẽ chọn ra một tầng đặc trưng theo tỷ lệ thích hợp nhất để lấy ra các vùng đặc trưng đề xuất. Công thức mà ta chọn được thể hiện như sau:

Trong đó , và là đặc trưng tầng của FPN được dùng để lấy ra các vùng đặc trưng đề xuất. Do đó ví dụ như thì đặc trưng ở tầng sẽ được chọn. Sau đó ta sẽ áp dụng RoI pooling và đưa kết quả vào mạng Fast R – CNN để kết thúc việc dự đoán.

**2.4.3. Double Head**

Cấu trúc hai lớp (một lớp fully connected và một lớp tích chập) được sử dụng rất nhiều trong các phương pháp phát hiện đối tượng dựa trên R – CNN cho hai công việc: phân loại hộp đề xuất và hồi quy tọa độ. Tuy nhiên trong nghiên cứu của mình, Wu cùng cộng sự [22] cho rằng có một sự thiếu hiểu biết nhất định rằng làm thế nào mà hai lớp này có thể hoạt động cho cả hai công việc trên. Các kết quả phân tích của các tác giả cho thấy lớp fully connected phù hợp hơn cho việc phân loại hộp đề xuất hơn, và lớp tích chập lại phù hợp cho việc hồi quy tọa độ hơn. Phân tích thêm đầu ra của hai lớp này, các tác giả nhận thấy đầu ra của fully connected nhạy cảm hơn về không gian so với lớp tích chập. Do đó các tác giả đã đề xuất phương pháp Double-Head: sử dụng một lớp fully connected cho việc phân loại và một lớp tích chập cho việc hồi quy hộp bao.

Diagram

Description automatically generated

**Hình Framework DoubleHead**

**2.4.4. GroIE**

Trong các phương pháp phát hiện đối tượng 2 bước như Faster R – CNN, lớp trích xuất vùng đề xuất (Region of Interest layer) đóng vai trò quan trọng. Cụ thể, nó dùng để trích xuất một tập hợp con nhất quán của các đặc trưng từ một lớp mạng FPN được đặt ở đầu kiến trúc.

Nhận thấy được các lớp RoI trước đây chỉ chọn một lwps tốt nhất từ FPN là một hạn chế, Leonardo và cộng sự dã đề xuất Generic RoI Extractor (GRoIE) [23], giới thiệu các khối xây dựng phi cục bộ (non-local building blocks) và cơ chế chú ý (attention mechanism) để tăng hiệu suất.

Diagram

Description automatically generated

**Hình Bốn module trong GRoIE**

**RoI pooler module:**  đây là module thực hiện max pooling trên các vùng đề xuất không đồng nhất để có được các biểu diễn có kích thước cố định. Các tác giả sử dụng kỹ thuật RoI Align vì nó làm giảm một vùng đặc trưng hình chữ nhật bằng cách chia RoI ban đầu trong các hộp bằng nhau và áp dụng nội suy song tuyến bên trong mỗi hộp, điều này giúp tránh lượng tử hóa pixel.

**Pre-processing module**: Mục tiêu của module này là áp dụng một bản tổng hợp sơ bộ cho các vùng được gộp (pooled). Điều này mang lại cho mạng một mức độ tự do bổ sung dành riêng cho từng tỷ lệ ảnh. Module này được dùng để tiền xử lý các đặc trưng và thường được áp dụng bởi một lớp tích chập được liên kết với mỗi tỷ lệ hình ảnh.

**Aggregation module**: module này định nghĩa bằng cách nào có thể tổng hợp các RoI đơn đến từ mỗi nhánh. Toán tử thường được dùng nhất là nối (concatenation) và tính tổng (summation). Trong nghiên cứu này các tác giả thấy rằng toán tử tính tổng có khả năng tối thiểu hóa số lượn đặc trưng cần tính cho các lớp tiếp theo, và điều này yêu cầu ít nỗ lực hơn từ mạng để hội tụ, hướng đến ổn định trong quá trình huấn luyện.

**Post-processing module:** Đây là một bước hậu xử lý bổ sung áp dụng cho các đặc trưng đã được hợp nhất trước khi trả về. Nó cho phép mạng tìm hiểu các đặc trưng toàn cục cùng xem xét tất cả kích thước. Để tăng cường sức mạnh thông tin của RoI cuối cùng, ba loại module được cân nhắc cho hậu xử lý: một lớp tích chập, một lớp phi cục bộ và một lớp chú ý (attention layer).

Tổng hợp lại, bắt đầu từ một vùng đề xuất bởi RPN, đối với mỗi tỷ lệ, một RoI cố định kích thước sẽ được gộp (pool) từ vùng này. Sau đó kết quả thu được là đặc trưng. Các đặc trưng này đầu tiên được tiền xử lý một cách riêng biệt, sau đó được gộp thành một đặc trưng duy nhất. Cuối cùng, hậu xử lý sẽ được áp dụng để trích xuất các thông tin toàn cục. Kiến trúc này cho phép đóng góp ngang nhau ở mỗi tỷ lệ và được hưởng lợi từ thông tin có trong tất cả các lớp FPN bằng cách khắc phục những hạn chế vốn có trong việc lựa chọn tùy ý một lớp FPN duy nhất. Đáng lưu ý rằng phương pháp này áp dụng cả cho phát hiện đối tượng và phân đoạn đối tượng.

**2.4.5. Guided Anchoring**

Guided Anchoring [19] là một phương pháp được đề xuất để cải thiện quá trình tạo ra các region proposal hiệu quả và linh hoạt hơn, do đó baseline của phương pháp này là Region Proposal Network (RPN). Phương pháp này sẽ dự đoán vị trí một điểm trung tâm của vật thể mà có khả năng tồn tại cũng như các scales và tỷ lệ khung hình tại các vị trí khác nhau. Điểm đặc biệt ở đây là Guided Anchoring hoàn toàn bỏ đi cơ chế khởi tạo các hộp neo (anchor) với các thông số mặc định, nói cách khác, kích thước và tỉ lệ khung hình của một anchor bây giờ có thể thay đổi một cách linh hoạt thay vì cố định như trước, từ đó tạo ra sự đa dạng trong quá trình học kèm theo phát triển khả năng phát hiện những vật thể có kích thước đặc biệt. Hơn nữa, tác giả còn nghiên cứu tầm ảnh hưởng của các vùng đề xuất chất lượng cao (high-quality proposal) trong các two-stage detectors.

Theo nghiên cứu, các cơ chế anchor generation đi trước được xây dựng trên nguyên lý tạo ra một loạt các anchor rất dày đặc, từ đó dẫn đến tình trạng xuất hiện không ít anchors không bọc lấy vật thể, nói cách khác là thừa thãi. Tuy nhiên, với Guided Anchoring thì các vấn đề sai lệch (misalignment) và mâu thuẫn (inconsistency) sẽ được nghiên cứu và giải quyết triệt để.

Trong Guided Anchoring, tác giả sử dụng một mô đun để tiến hành khởi tạo các anchor. Mô đun này sẽ có hai nhánh để dự đoán lần lượt vị trí trung tâm và kích thước của anchor như trên hình 3-3. Đối với nhánh Anchor Location Prediction, nơi mà sẽ dự đoán vị trí của anchor, nhánh này sẽ tạo ra một ánh xạ xác suất (probability map) để có thể tìm ra vị trí có khả năng tồn tại vật thể trong ảnh. Để có được một probability map thì tác giả đã sử dụng một mạng con (sub-network). Mạng này sẽ sử dụng 1x1 convolution trên feature map đầu vào để thu được một ánh xạ thể hiện objectness scores. Sau đó thực hiện chuyển đổi các objectness scores này thành giá trị xác suất thông qua một hàm element-wise sigmoid. Với một ngưỡng nhất định đã được thiết lập từ trước, họ sẽ nhờ vào ngưỡng này mà xác định ra những vùng tích cực (active region) mà có khả năng vật thể tồn tại.

Diagram

Description automatically generated

***Hình Framework của cơ bản Guided Anchoring.***

Đối với nhánh Anchor Shape Prediction, với tọa độ của vị trí đã được khởi tạo trước đó, nói cách khác là dựa vào đầu ra của nhánh Anchor Location Prediction để tiến hành dự đoán lần lượt tượng trưng cho chiều rộng và chiều cao của anchor, từ đó tìm ra được hình dạng anchor gần khớp với ground-truth bbox nhất. Tuy nhiên quá trình thực nghiệm cho thấy rằng việc dự đoán trực tiếp hai giá trị này không khả quan. Do đó, tác giả đã đưa ra những phép biến đổi nhất định để giải quyết vấn đề này bằng cách sử dụng một sub-network để thực hiện convolution để tạo ra một ánh xạ phù hợp.

Cần phải nhấn mạnh rằng, thiết kế này hoàn toàn khác xa với các cơ chế khởi tạo anchor thông thường vì trong Guided Anchoring, mỗi vị trí chỉ liên kết với một anchor được dự đoán kèm theo kích thước đa dạng thay vì liên kết với một bộ các anchor có kích thước được định nghĩa sẵn. Để khai thác và trích xuất các đặc trưng, tác giả đã nghĩ ra một cấu tạo của anchor-guided feature adaptation, thứ sẽ biến đổi các đặc trưng tại một vị trí cụ thể dựa trên các anchor shape đã được khởi tạo ở giai đoạn trước đó.

Trước hết, tiến hành dự đoán vùng offset dựa trên đầu ra của nhánh anchor-shape prediction và sau đó áp dụng deformable convolution để thu được các đặc trưng cần thiết cho bài toán classification và hồi quy bounding box. Trong quá trình nghiên cứu, các vùng đề xuất của GA-RPN đã có những điểm vượt trội với các vùng đề xuất của RPN thông thường. Đầu tiên, GA-RPN tạo ra số lượng các vùng đề xuất positive nhiều hơn. Không chỉ nhiều hơn, tỉ lệ chính xác của các high-IOU proposal của GA-RPN cũng cao hơn so với RPN thông thường. Qua đó việc thay thế RPN trong những phương pháp sẵn có bằng GA-RPN kết hợp với một chút tinh chỉnh đã mang lại những sự cải thiện đáng kể trong kết quả.

**2.4.6. Side-aware Boundary Localization (SABL)**

Các phương pháp phát hiện đối tượng hiện tại phụ thuộc vào bước hồi quy hộ bao (bounding box regresson) cho việc định vị đối tượng. Mặc đã có các quá trình cố gắng cải thiện trong những năm gần đây, độ chính xác của hồi quy hộp bao vẫn chưa được thỏa mãn, dẫn đến đây là hạn chế của phát hiện đối tượng. Wang cùng cộng sự nhận thấy rằng các cách tiếp cận trước đây chỉ chú ý vào dự đoán tọa độ trung tâm và kích thước (x, y, w, h), điều này không phải là một cách hiệu quả để thực hiện hồi quy hộp bao, đặc biệt là khi tồn tại các chuyển vị và phương sai lớn giữa các hộp mỏ neo và ground truth. Do đó các tác giả đã đề xuất phương pháp Side-Aware Boundary Localization (SABL) [24], trong đó mỗi bên của hộp bao sẽ được định vị lần lượt với một nhánh mạng chuyên dụng.

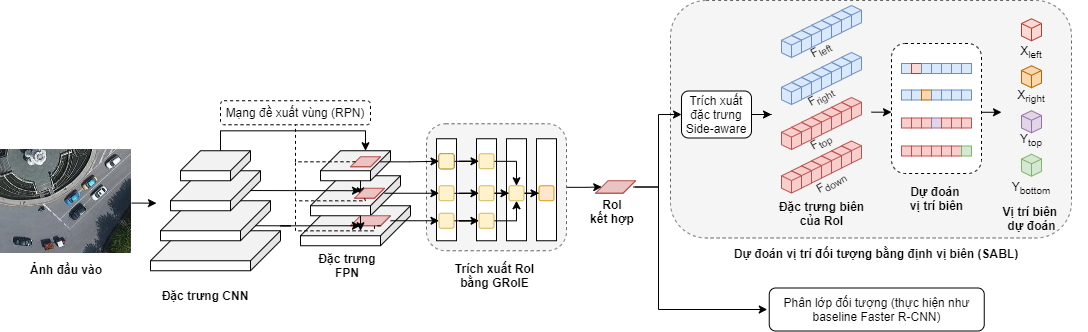
Diagram

Description automatically generated

**Hình Framework Side-Aware Boundary Localization**

Như ở Hình 2.24, đầu tiên phương pháp này sẽ trích xuất các đặc trưng ngang và dọc ( và ) bằng cách kết hợp các đặc trưng RoI dọc theo trục X và trục Y một cách lần lượt. sẽ được chia thành các đặc trưng side-aware . Sau đó ở mỗi bên của hộp bao, SABL đầu tiên chia các không gian đích thành nhiều nhóm và tìm kiếm vùng chứa biên thông qua việc tận dụng các đặc trưng side-aware. Nó sẽ tinh chỉnh lại các tọa độ biên bằng cách dự đoán thêm offset của chúng. Một framework chia nhóm hai bước như vậy sẽ giảm phương sai hồi quy và giảm bớt những khó khăn trong dự đoán. Hơn nữa, độ tin cậy của các nhóm ước tính cũng có thể giúp điều chỉnh điểm phân loại (classification score) và cải thiện hơn nữa hiệu xuất. Không những áp dụng được cho các phương pháp hai bước, nó còn có thể áp dụng cho các phương pháp phát hiện một bước.

**2.4.7. PP Đề xuất: Kết hợp GDSNet (GRoIE + Double Head + SABL)**

****