ĐẠI HỌC QUỐC GIA TP. HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN CUỐI KÌ

PHÁT HIỆN ĐỐI TƯỢNG TRÊN TÀI LIỆU DẠNG ẢNH TIẾNG VIỆT

Giảng viên bộ môn: TS. Mai Tiến Dũng

Khoa: Khoa học máy tính

Nhóm thực hiện:

1. **Nguyễn Nhật Trường 20522087**
2. **Lê Trương Ngọc Hải 20520481**

TP. HỒ CHÍ MINH, 2023

MỤC MỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Ý nghĩa
01	AP	Average Precision
02	mAP	Mean Average Precision
03	YOLO	You only look once
04	CNN	Convolutional Neural Networks
05	R-CNN	Region-based Convolutional Network
06	YOLOF	You Only Look One-level Feature
07	CEM	Crossline Extraction Module
08	DNN	Deep Neural Networks
09	DOD	Document Object Detection
10	SSD	Single Shot MultiBox Detector
11	RPN	Region Proposal Network
12	FD	False Discovery
13	IoU	Intersection over Union
14	CSP	Cross-stage-partial-connection
15	NMS	Non Maximum Suppression
16	FCOS	Fully Convolutional One-Stage Object Detection
17	GA-RPN	Region Proposal by Guided Anchoring
18	FPN	Feature Pyramid Network
19	FSAF	Feature Selective Anchor-Free
20	FPS	Frame per second
21	FC-head	Fully-connected head
22	CONV-head	Convolution head

Chương 1. **Tổng quan đồ án**

1.1. Lí do chọn đồ án

Nhu cầu trích xuất thông tin từ tài liệu kỹ thuật số tăng nhanh chóng do sự phát triển của công nghệ và nhu cầu số hóa tài liệu. Tuy nhiên bố cục và cách trình bày của văn bản mang đến một số khó khăn cho vấn đề này. Nhận thấy tiềm năng của bài toán phát hiện đối tượng trên trang tài liệu và sự thiếu hụt của các mô hình nhận biết đối tượng trên tài liệu các ngôn ngữ ít phổ biến, chúng em quyết định tập trung vào nghiên cứu bài toán phát hiện đối tượng tài liệu tiếng Việt này.

1.2. Giới thiệu bài toán

Quá trình số hóa tài liệu đã và đang diễn ra trong nhiều tổ chức và doanh nghiệp kể từ khi thời đại công nghệ 4.0 bắt đầu phát triển mạnh mẽ, các tài liệu truyền thống như (giấy, sổ, hóa đơn) đang dần chuyển hóa và thay bằng các tài liệu số (PDF, WORD, EXCEL) được lưu trữ trên các dịch vụ điện toán đám mây để thuận tiện cho truy cập, tìm kiếm, lưu trữ tài liệu. Với lượng lớn tài liệu như vậy việc tìm kiếm tài liệu trở nên khó khăn hơn bao giờ hết. Thế nên một mô hình tốt để nhận diện các thành phần, đối tượng có trong ảnh tài liệu là thật sự cần thiết. Từ đó, chúng tôi quyết định đặt vấn đề vào góc nhìn của bài toán object detection. Document Object Detection (DOD) hướng đến nhiệm vụ phát hiện tự động các thành phần quan trọng (Caption, Table, Figure, …) và cấu trúc của trang tài liệu. Bởi sự phổ biến và thông dụng của ngôn ngữ Anh, Trung, hầu hết những mô hình phát hiện đối tượng dành cho tài liệu hiện nay thường tập trung vào hai ngôn ngữ này. Tuy nhiên, các ngôn ngữ khác nhau cũng có các đặc trưng khác nhau từ cách trình bày khác nhau cho các đối tượng như chú thích, công thức cho đến ngữ nghĩa và ngữ pháp trong tài liệu.

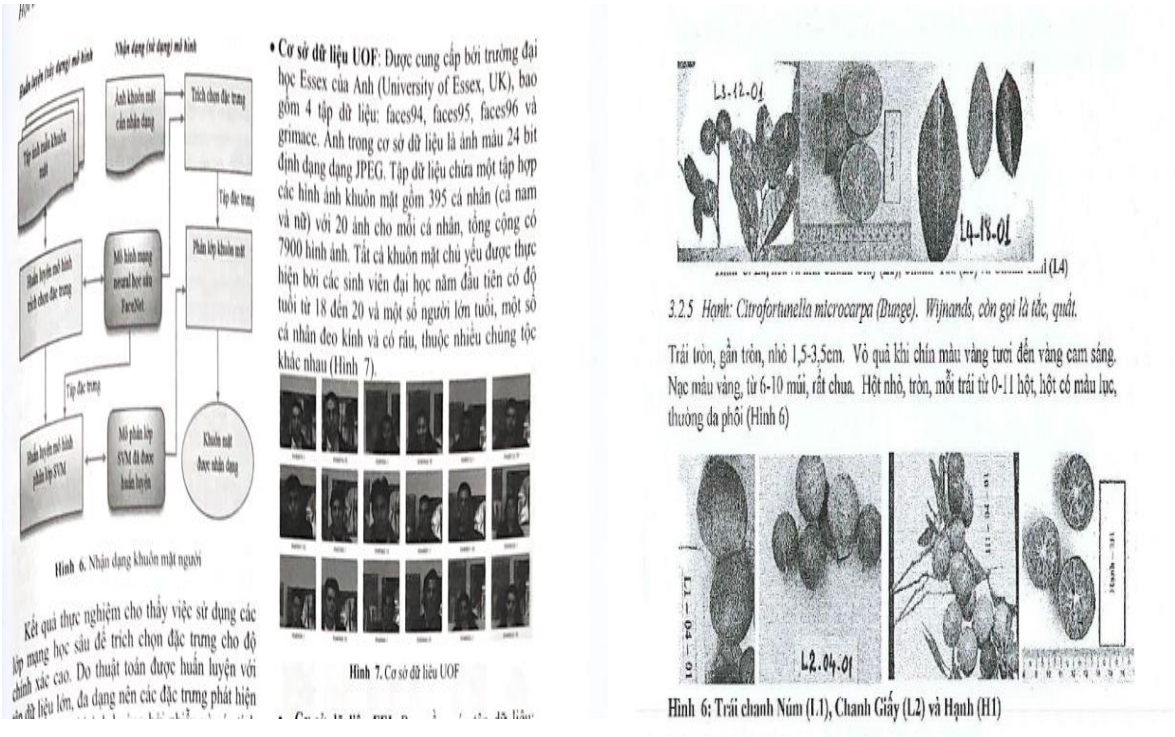
Trong nghiên cứu này, chúng tôi tiến hành thử nghiệm các phương pháp state-of-the-art như: Double-Head RCNN, Libra RCNN, Guided Anchoring trên UIT-DODV dataset, bộ dataset tiếng Việt đầu tiên với các đối tượng của hình ảnh đầu vào là Caption, Table, Figure và Formula, và đồng thời cũng đề xuất một phương pháp qua việc tận dụng điểm mạnh các các phương pháp có trước và kết hợp chúng lại, hướng tới kết quả tốt nhất trong bài toán phát hiện đối tượng. Về bộ dữ liệu UIT-DODV, đây là một sự tổng hợp các tài liệu Tiếng Việt, một ngôn ngữ khá mới mẻ trong nghiên cứu, do đó mang lại nhiều thách thức mới, ví dụ như cách trình bày các đối tượng ngữ nghĩa tạo ra nhiều khó khăn trong việc rút trích đặc trưng các thông tin, các công thức không chỉ là công thức toán học bình thường mà còn ở các dạng không thuộc toán học, thậm chí đến các ký hiệu, ký tự cũng rất khác biệt.

Thông qua quá trình thực nghiệm, chúng tôi đã đưa ra những đánh giá khách quan của những mô hình phát hiện đối tượng phổ biến trên bộ dữ liệu UIT-DODV. Cụ thể, chúng tôi đánh giá kết quả trên các phương pháp sau: Double-Head-R-CNN, Libra R-CNN, Guided Anchoring. Chúng tôi đã đạt được kết quả ban đầu với Guided Anchoring đạt 73.6% trên độ đo mAP. Dựa vào những phân tích từ hiệu suất của các mô hình phát hiện đối tượng đã đề cập, chúng tôi đã đề xuất mô hình phát hiện đối tượng dựa trên hai phương pháp là Cascade R-CNN và Guided Anchoring. Mô hình đề xuất của chúng tôi đã đạt được mAP lên đến 76.6%, cao hơn baseline được trình bày đến 2.1%.

1.3. Thách thức bài toán

Trong bài toán phát hiện đối tượng trên tài liệu dạng ảnh, bảng là hai đối tượng chủ yếu và thường bắt gặp nhất trong quá trình nghiên cứu. Đối tượng "Bảng" thường có kết cấu rất gọn và những đặc trưng cơ bản gồm khung và các đường thẳng, sử dụng cho việc thống kê, tóm tắt nội dung muốn thể hiện trong nhiều lĩnh vực từ công việc tài chính cho đến hoạt động nghiên cứu. Do đó, bởi sự phổ biến của đối tượng này, hầu hết các nghiên cứu phát hiện đối tượng tài liệu đều nhắm tới.

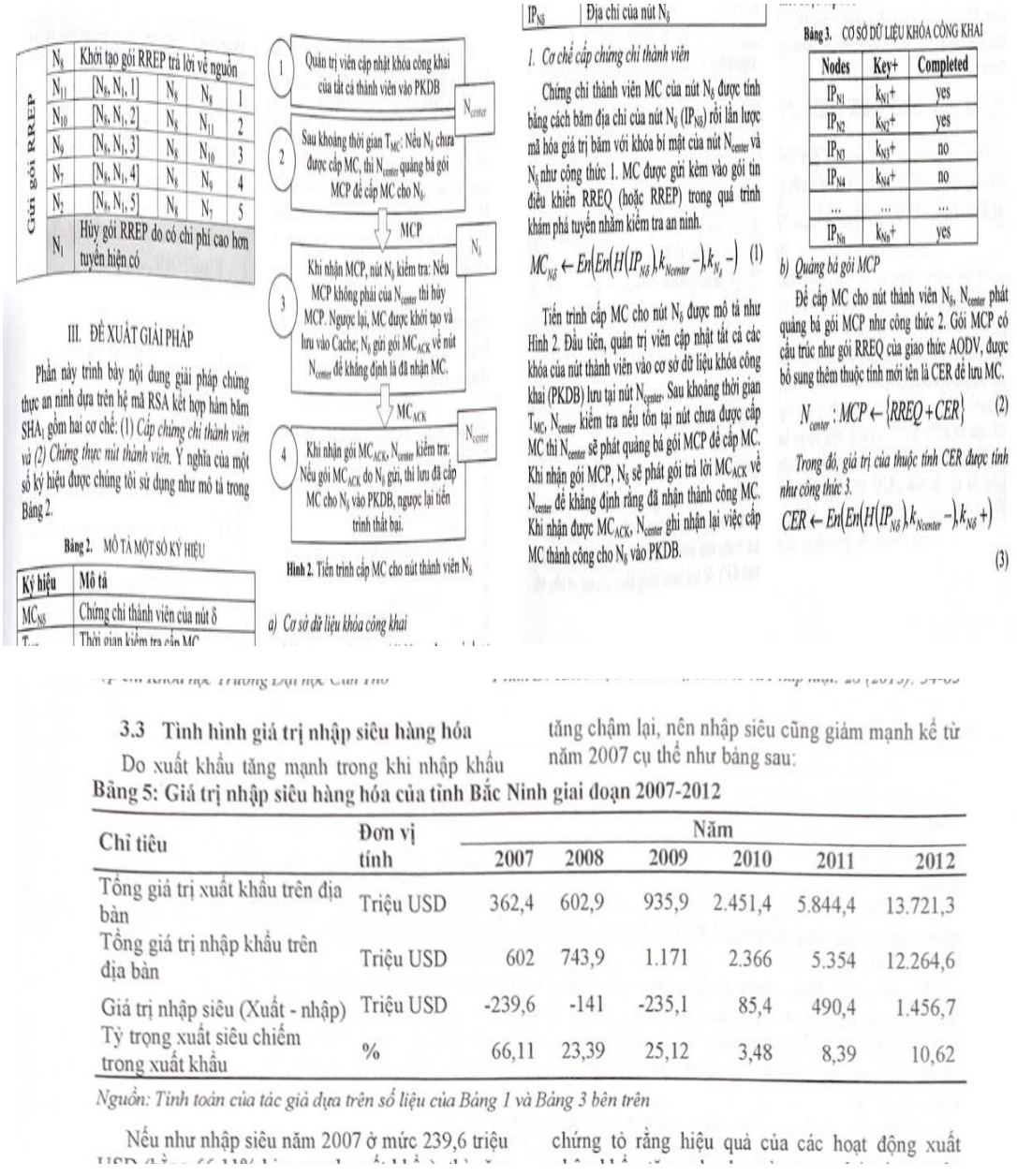
Tương tự với bảng, đối tượng "Ảnh" cũng rất phổ biến khi các nhà nghiên cứu tập trung nghiên cứu rất kĩ đối tượng này. Như ta biết, ảnh được dùng để minh họa nội dung mà ta muốn truyền tải, tóm gọn lý thuyết và trực quan một cách khách quan dữ liệu như các ví dụ trong hình 1-1. Hơn nữa, lượng thông tin và đặc trưng ở các ảnh là rất dồi dào và dễ dàng rút trích, nên có thể xem đây là một đặc điểm có thể tận dụng để có thể mang lại kết quả tốt nhất cho bài toán DOD.



Hình 1 - 1: *Một số thể hiện của đối tượng "Ảnh" trong bộ dữ liệu UIT-DODV*

Tuy nhiên, đối với các dữ liệu mang tính tài liệu, các đối tượng chi tiết như "Công thức" hay "Tiêu đề" lại tạo ra rất nhiều khó khăn trong việc rút trích thông tin bởi chúng có sự tương đồng với nhau, hơn nữa, các đặc trưng rất nhỏ và khó phân biệt gây ảnh hưởng tới hiệu quả của mô hình. Do hai đối tượng này được cấu thành từ chữ cái, ký hiệu,… nên rất có khả năng bị phân loại, phát hiện nhầm và thậm chí bỏ sót vì quá tương đồng với background. Bên cạnh đó, lượng đặc trưng khó rút trích và những đặc điểm khó nhận dạng tạo ra không ít khó khăn cho quá trình nhận diện

đối tượng trong ảnh, tác động không nhỏ tới hiệu suất và tốc độ của quá trình huấn luyện trong thực nghiệm.



Hình 1 - 2*: Một số thể hiện về kết cấu của đối tượng "Bảng" trong bộ dữ liệu UIT-DODV*

Tuy bảng và hình ảnh là các đối tượng được đánh giá là hiệu quả nhất trong các nghiên cứu nhưng các kích thước, độ phân giải và hình dạng khác nhau gây không ít thách thức cho các mô hình. Hơn nữa, kiến trúc của một bảng là không cố định khi

người khác có thể xây dựng bảng dưới góc nhìn của chính họ, nói cách khác, bảng có thể không có viền, không có chia cột hoặc có một số bảng thì không chia cả hàng như trong hình 1-2. Chính sự khác nhau này cũng tạo ra sự đa dạng về kết cấu và thành phần của dữ liệu, thứ mà sẽ được mô hình "học" trong quá trình thực nghiệm. Tuy nhiên, sự đa dạng này gây không ít khó khăn cho quá trình xây dựng và huấn luyện nhưng nhờ vậy, mô hình sẽ dần được hoàn thiện và cải thiện tốt nhất có thể.

Không chỉ có những thách thức về dữ liệu, mà còn là những giới hạn nhất định về chi phí, điều kiện nghiên cứu. Việc chọn một mô hình Deep Learning thích hợp với khả năng đáp ứng là một điều đáng cân nhắc, đồng thời, mô hình đề xuất phải thể hiện được điểm mạnh của nó một cách mới mẻ, sáng tạo.

1.4. Cấu trúc đồ án

Chương 1: Tổng quan đồ án. Lí do chọn đồ án, cung cấp cái nhìn tổng quan về đồ án và các thách thức trong quá trình thực nghiệm thu được.

Chương 2: Các công trình nghiên cứu liên quan. Các nền tảng kiến thức về các phương pháp one-stage và two-stage sẵn có trong bài toán phát hiện đối tượng.

Chương 3: Phương pháp đề xuất cải thiện hiệu xuất của RCNN. Trình bày về các phương pháp cải thiện R-CNN và đề xuất phương pháp mới.

Chương 4: Thực nghiệm – Đánh giá. Trình bày quá trình cài đặt thực nghiệm, các thông số chi tiết, phương pháp đánh giá và phân tích kết quả.

Chương 5: Kết luận và hướng phát triển. Đưa ra tổng kết đáng chú ý đạt được trong đề tài và trình bày hướng phát triển trong tương lai.

Chương 2. **Các công trình nghiên cứu liên quan**

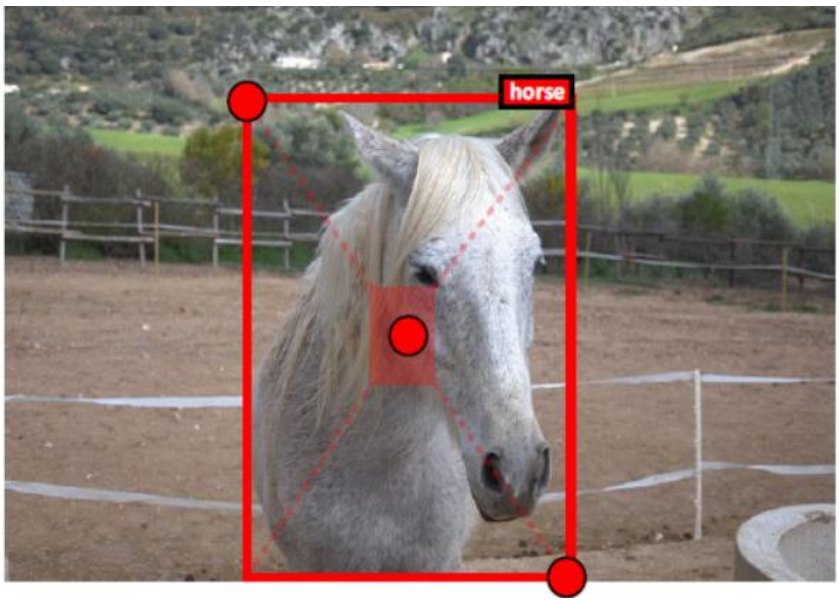
2.1. Các phương pháp one-stage

2.1.1. Giới thiệu phương pháp one-stage

Một các tiếp cận khá phổ biến hiện nay như one-stage object detection với 1 số model điển hình như: SSD, YOLO v2.3, RetinaNet. Các mô hình one-stage thường nhanh hơn tuy nhiên độ chính xác của mô hình thường kém hơn so với các phương pháp two-stage object detection. Nhưng cũng có một số mô hình tỏ ra vượt trội hơn two-stage như RetinaNet.

2.1.2. CenterNet

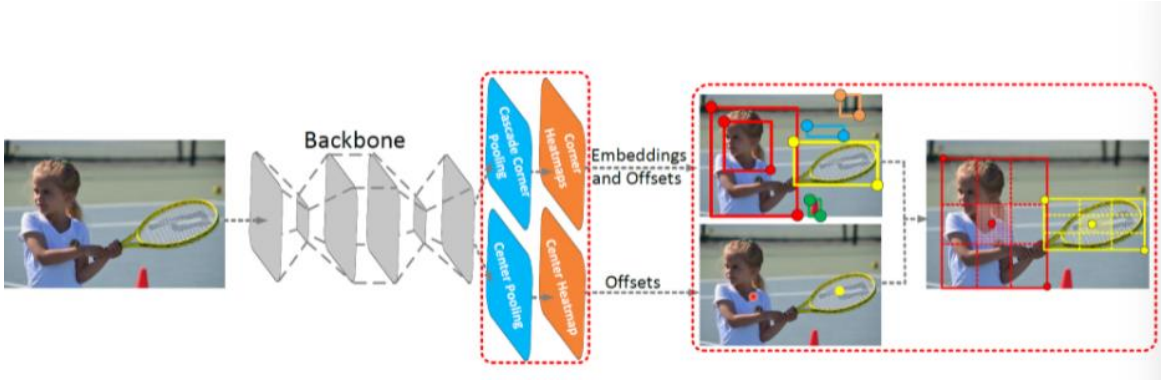
Duan et al. [6] đề xuất CenterNet là dựa trên hướng tiếp cận đưa bài toán phát hiện đối tượng (object detection) về bài toán ước tính điểm chính (keypoint estimation), từ đó suy ra kích thước và tính toán được bounding box cho bài toán phát hiện vật thể.



Hình 2 - 1: Bộ ba điểm phát hiện đối tượng trong CenterNet [6].

Ở phương pháp CornerNet còn khá nhiều hạn chế vì dễ bị nhầm với các đặc trưng cạnh nếu có trong ảnh đầu vào. Bên cạnh đó, việc xác định hai keypoint ở hai góc

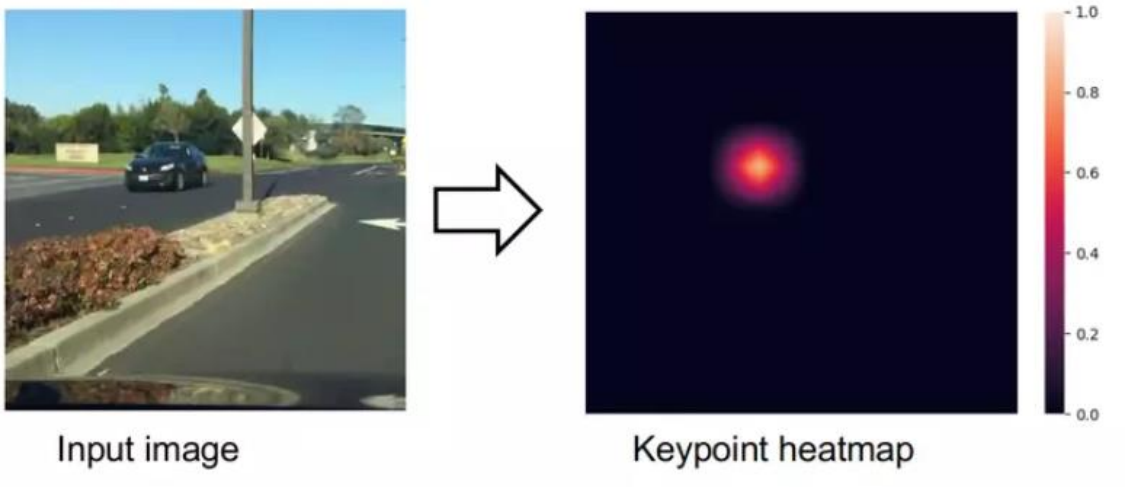
có cùng một đối tượng hay không cũng gặp một số vấn đề. Do khả năng tiếp nhận các thông tin toàn cục và keypoint góc thường nằm bên ngoài đối tượng do đó rất khó khăn để nhóm hai góc lại thành cùng một đối tượng. Để giải quyết vấn đề này thì CenterNet đã ra đời. CenterNet nhận dạng mỗi đối tượng bằng bộ ba điểm chính thay vì sử dụng cặp điểm như là CornerNet điều đó giúp cho CenterNet cải thiện được precision và recall. Ngoài ra, CenterNet đã học thêm được những thông tin, dữ liệu toàn cục và nhờ vào keypoint cho điểm trung tâm của đối tượng nên giúp cho CenterNet cải thiện hiệu quả khi phát hiện đối tượng.



Hình 2 - 2: Kiến trúc mạng CenterNet.[1]

CenterNet sử dụng baseline là kiến trúc CornerNet và đi qua một backbone CNN trích xuất các đặc trưng từ ảnh, sau đó sử dụng hai lớp pooling để sinh ra các heatmap cho corner keypoint và center keypoint. Hai lớp đó là Cascade Corner Pooling và Center Pooling và chính hai lớp này giúp mô hình cải thiện về cả độ chính xác và FD (false discovery). Nếu như AP(Average Precision) là giá trị đánh giá độ chính xác của mô hình xác định đối tượng như ở SSD, Faster RCNN,… ở IoU = [0.05;0.05;0.5] trên dataset MS-COCO,…Thì ngược lại FD là giá trị đo lường số lượng bounding box không chính xác nói cách khác là có tỉ lệ IoU dưới ngưỡng cho trước.

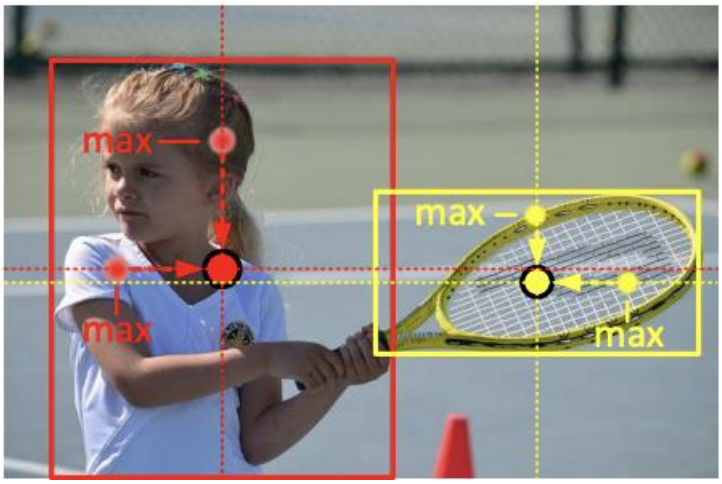
[1] https://medium.com/nerd-for-tech/review-centernet-keypoint-triplets-for-object-detection-object-detection-26feee780efc



Hình 2 - 3: Hình minh họa bản đồ nhiệt (Heatmap)[2]

Mỗi điểm trên **heatmap** (bản đồ nhiệt) tương ứng với một keypoint cùng với score là xác xuất keypoint là tâm của đối tượng. Điểm giữa của một đối tượng không thể bao gồm toàn bộ đặc điểm nhận dạng của đối tượng đó. Giả sử một vấn đề đặt ra điểm giữa của con người thường nằm ở phần thân trong khi phần đầu là phần quan trọng nhất giúp ta nhận ra đối tượng.

2.1.2.1 CenterPooling



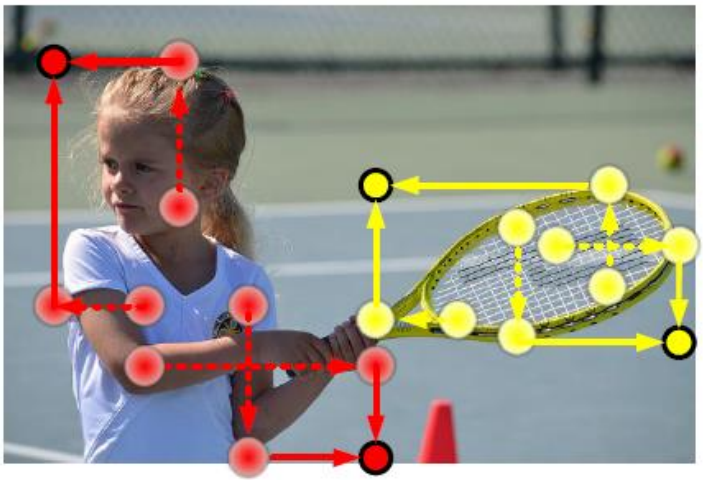
Hình 2 - 4: Hình ảnh minh họa Center Pooling[3]

[2] https://itzone.com.vn/vi/article/centernet-keypoint-triplets-for-object-detection-huong-di-moi-trong-bai-toan-object-detection/

Để giải quyết vấn đề trên, tác giả đề xuất **Center Pooling** giúp mô hình có thể học nhiều thông tin hơn cho toàn bộ đối tượng. Lớp pooling này sẽ nhận đầu vào là một feature map được trích xuất qua một mạng CNN, sau đó tìm giá trị lớn nhất theo chiều ngang và chiều dọc và cộng chúng lại với nhau nhờ việc này mà mô hình có thêm đặc trưng của cả đối tượng.

Center Pooling được thực hiện gồm các bước: ban đầu sẽ tạo ra k bounding box bằng thuật toán sử dụng trong CornerNet, sau đó chọn k center keypoint có score xác xuất cao nhất và sử dụng các offset tương ứng với các keypoint đó để xác định các center keypoint đó trên ảnh đầu vào. Kế đến là xác định vùng nằm giữa mỗi bounding box và kiểm tra nếu center keypoint nằm ở trong vùng trung tâm thì giữ nguyên bounding box, còn nếu không thì xóa. Độ chính xác của mỗi bounding box này bằng trung bình độ chính xác của ba keypoint xác định nó.

2.1.2.2 Cascade Corner Pooling

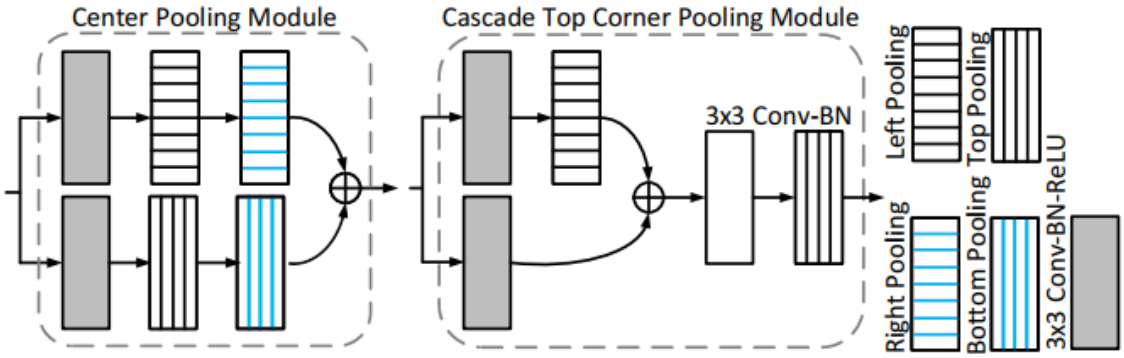


Hình 2 - 5: Hình ảnh minh họa cascade corner pooling

Cascade Corner Pooling sinh ra để khắc phục và cải thiện việc học các thông tin toàn cục của Corner Pooling trong CornerNet. Lớp pooling này tìm các keypoint

[3]https://itzone.com.vn/vi/article/centernet-keypoint-triplets-for-object-detection-huong-di-moi-trong-bai-toan-object-detection/

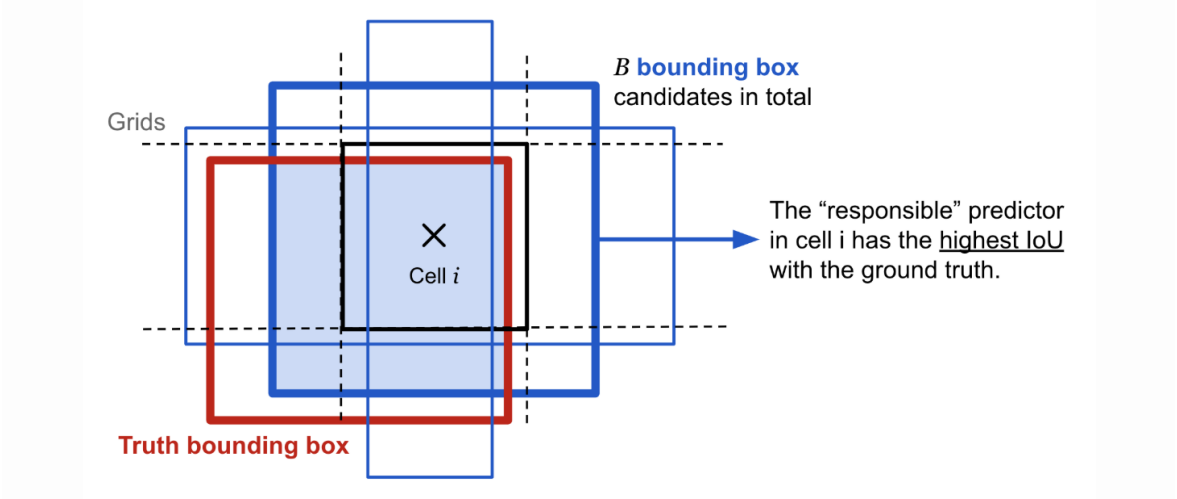
góc bằng các tìm giá trị lớn nhất trên một boundary (đường viền). Sau đó dọc theo giá trị lớn nhất đó nhìn vào trong đối tượng để tìm internal maximum value (giá trị lớn nhất nội bộ), sau đó cộng hai giá trị lớn nhất này lại với nhau. Như trên hình 2-5 ta có thể thấy nếu đang xét ở topmost boundary thì dọc xuống bottom, ở leftmost boundary thì dọc sang phải. Nhờ cách này mà góc chứa cả thông tin trên boundary và thông tin của đối tượng.



Hình 2 - 6: Cấu trúc của center pooling module và cascade top corner pooling module

2.1.3. Anchor-based Detectors

Các phương pháp phát hiện đối tượng nổi tiếng ngày nay như Fast-RCNN, YOLOv3, SSD, RetinaNet, … đều dựa trên cơ chế được gọi là khởi tạo hộp neo (anchor generation) hay còn gọi là các pre-define boxes với mục đích dự đoán vị trí của các bounding box của vật thể dựa trên các anchor đó. Tuy nhiên, việc sử dụng các anchor tạo ra quá nhiều siêu tham số và hao tốn tài nguyên rất đáng kể. Đáng nói hơn, là các siêu tham số có sự tác động trực tiếp đến kết quả cuối cùng dù cho những thay đổi là tối thiểu. Do đó, kết quả mô hình bị chi phối hoàn toàn ở giai đoạn anchor generation và chất lượng của các anchor thu được (sử dụng cho quá trình huấn luyện sau này).



Hình 2 - 7: Hình ảnh mô tả sử dụng anchor-base trong YOLOv3[4]

2.1.4. Anchor-Free Detectors

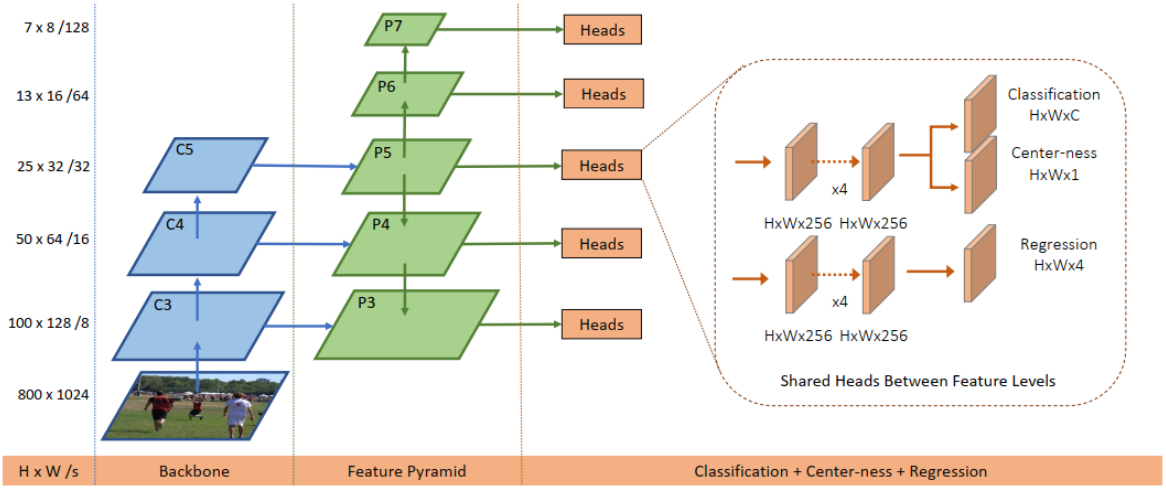
Một cách tiếp cận khác ở đây là phương pháp không dựa trên hộp neo (anchor-free method) tức là bỏ qua các anchor. Ngày càng có nhiều bộ phát hiện hoạt động trên nguyên lý anchor-free được sử dụng và vận hành cho bài toán cụ thể bởi các nhà khoa học đã đánh giá chúng tốt hơn hẳn so với các phương pháp anchor-based lúc đó. Những sự vượt trội được thể hiện ở hai phần rõ rệt: phần đánh giá (mAP) và thể hiện của model (FPS & Flops). Bộ phát hiện anchor-free phổ biến hiện nay có thể nhắc đến như DenseBox, FCOS/ FSAF, CornerNet, CenterNet, ...

2.1.5. FCOS

Hầu hết tất cả các bộ phát hiện đối tượng hiện nay như là RetinaNet, SSD, YOLOv3, và Faster RCNN đều dựa trên việc xác định trước các anchor boxes. Ngược lại thì ở phương phương này FCOS sử dụng anchor box free cũng như bỏ qua các vùng đề xuất. Thông qua việc bỏ các bộ pre-defined của anchor boxes do đó FCOS tránh được các tính toán phức tạp liên quan đến anchor boxes như là

[4] https://towardsdatascience.com/forget-the-hassles-of-anchor-boxes-with-fcos-fully-convolutional-one-stage-object-detection-fc0e25622e1c

overlapping trong quá trình huấn luyện. Quan trọng hơn là việc sử dụng free-anchor giúp tránh khỏi tất cả siêu tham số liên quan đến anchor boxes, cái mà rất sensitive trong thể hiện đối tượng cuối cùng.



Hình 2 - 8: Kiến trúc của FCOS

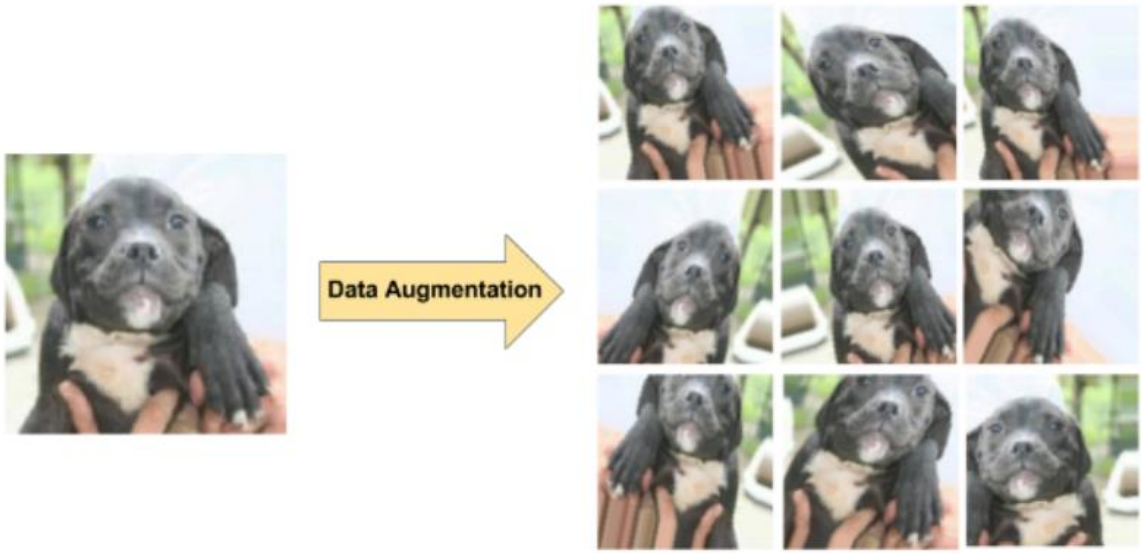
2.1.6. YOLOv4

YOLOv4 [20] là một loạt các cải tiến về tốc độ so với YOLOv3, đáp ứng được vấn đề tiêu tốn quá nhiều thời gian trong quá trình đạt đến sự hội tụ của các phiên bản trước. YOLOv4 là một sự kết hợp hiệu quả giữa kiến trúc CSPNet và một backbone CNN đã được pretrained với gần 1 triệu tấm ảnh trên bộ dữ liệu ImageNet, Darknet-53 (như trong YOLOv3). Hầu hết các mô hình chính xác hiện đại yêu cầu nhiều GPU để đào tạo với kích thước mimi-batch lớn và việc thực hiện điều này với một GPU khiến quá trình huấn luyện chậm và không thực tế. Với YOLOv4 sẽ giải quyết vấn đề này bằng cách tạo ra một bộ phát hiện đối tượng có thể được huấn luyện trên một GPU duy nhất với kích thước mini-batch nhỏ hơn. Điều này giúp cho việc đào tạo một bộ phát hiện đối tượng siêu nhanh và chính xác chỉ với duy nhất một GPU 1080 ti hoặc 2080 Ti.

YOLOv4 đạt được kết quả cao với tốc độ thời gian thực trên bộ dữ liệu MS COCO lên đến 43.5% AP với thời gian là 65 FPS trên một Tesla V100. Để đạt được kết

quả đó, các tác giả đã kết hợp một số đặc trăng như Weighted-Residual-Connections (WRC), Cross-Stage-Partial-connections(CSP), Cross mini-Batch Normalization (CmBN), Self-adversarial-training (SAT) and Mish-activation, Tăng cường dữ liệu Mosaic, DropBlock regularization, and CIoU loss. Chúng được gọi là các đặc trưng tổng thể vì chúng phải hoạt động tốt một cách độc lập với các tác vụ thị giác máy tính, bộ dữ liệu và các mô hình.

Bag of Freebies(BoF): Những phương pháp giúp cải thiện kết quả trong quá trình huấn luyện mà không làm ảnh hưởng tới tốc độ học của mô hình. Một vài phương pháp rất phổ biến có thể nhắc đến là tăng cường dữ liệu (data augmentation), mất cân bằng lớp(class imbalance), hàm chi phí(cost function), nhãn mềm(soft labeling). Một ví dụ về BOF là tăng cường dữ liệu (data augmentation) sẽ làm tăng khả năng tổng quát của mô hình. Để làm điều đó, họ sẽ thực hiện các biến dạng số liệu ảnh như: thay đổi độ sáng, độ bão hòa, độ tương phản và nhiễu hoặc họ có thể làm biến dạng hình học của một hình ảnh như xoay, cắt xén, ... Những kỹ thuật này là một ví dụ rõ về BoF và chúng giúp cải thiện độ chính xác cho bộ phát hiện đối tượng. Một điều cần chú ý là trong các bài toán phát hiện đối tượng thì bounding boxes cũng phải được áp dụng các phép biến đổi tương tự.



Hình 2 - 9: Ví dụ về data augmentation[5]

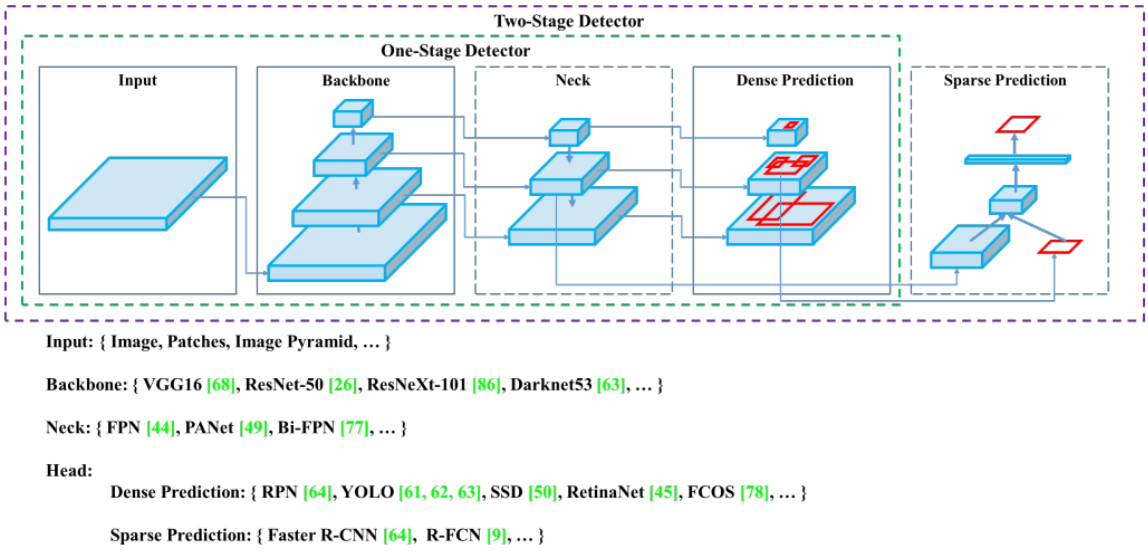
Bag of Specials(BoS): Những phương pháp hi sinh một chút tốc độ inference mà làm cải thiện độ chính xác của mô hình đáng kể. Những phương pháp này bao gồm tăng trường cục bộ (receptive field), điều hướng sự tập trung (attention), kết hợp các thông tin của feature maps với nhau (feature intergration) như skip-connection & FPN (Feature Paramyd Network), hậu xử lý như NMS (Non Maximum Suppression).

Mặc dù YOLO là bộ phát hiện one-stage, nhưng cũng có những bộ phát hiện two-stage như R-CNN, Fast R-CNN và Faster R-CNN những mô hình này có độ chính xác cao nhưng khá chậm. Chúng tôi sẽ đề cập đến mô hình object detection được cấu tạo bởi các thành phần:

- Input: Ảnh, Patches, Image Pyramid
- Backbone: Các mô hình như VGG16, ResNet-50, SpineNet, EffcientNet-B0/B7, CSPResNeXt50, CSPDarknet53,... được sử dụng làm trình rút trích đặc trưng. Chúng được huấn luyện trước về tập dữ liệu phân loại hình ảnh, như ImageNet, và sau đó được fine-tuned trên bộ dữ liệu phát hiện. Các mạng này sẽ tạo ra các cấp độ đặc trưng khác nhau với các ngữ nghĩa cao hơn khi mạng trở nên sâu hơn (nhiều layers hơn) và rất hữu ích cho các phần sau của mạng phát hiện đối tượng.
- Neck: Đây là những lớp bổ sung đi vào giữa backbone và head. Chúng được sử dụng để trích xuất các feature maps khác nhau của các giai đoạn khác nhau của backbone. Phần neck có thể được ví dụ như:
    - Khối bổ sung (Additional blocks): SPP, ASPP, RFB, SAM
    - Các khối tổng hợp đường dẫn (Path-aggregation): FPN, PAN, NAS-FPN, Fully-connected FPN, BiFPN, ASFF, SFAM

[5] https://towardsdatascience.com/yolo-v4-optimal-speed-accuracy-for-object-detection-79896ed47b50

- Heads: Đây là một mạng chịu trách nhiệm thực hiện phần phân loại và hồi quy các bounding box. Đầu ra duy nhất có thể trong giống như (tùy thuộc vào việc triển khai): 4 giá trị mô tả bounding box (x, y, w, h) và xác xuất của k lớp + 1 (thêm một lớp cho background). Các bộ phát hiện đối tượng dựa trên anchor-based, như YOLO sẽ áp dụng mạng head cho mỗi anchor box. Các bộ phát hiện đối tượng phổ biến khác như:
    - Dự đoán dày đặc – Dense Prediction (one-stage):
        - RPN, SSD, YOLO, RetinaNet (anchor-based)
        - CornerNet, CenterNet, MatrixNet, FCOS (anchor-free)
    - Dự đoán thưa thớt – Sparse Prediction (two-stage):
        - Faster R-CNN, R-FCN, Mask R-CNN (anchor-based)
        - RepPoints (anchor-free)

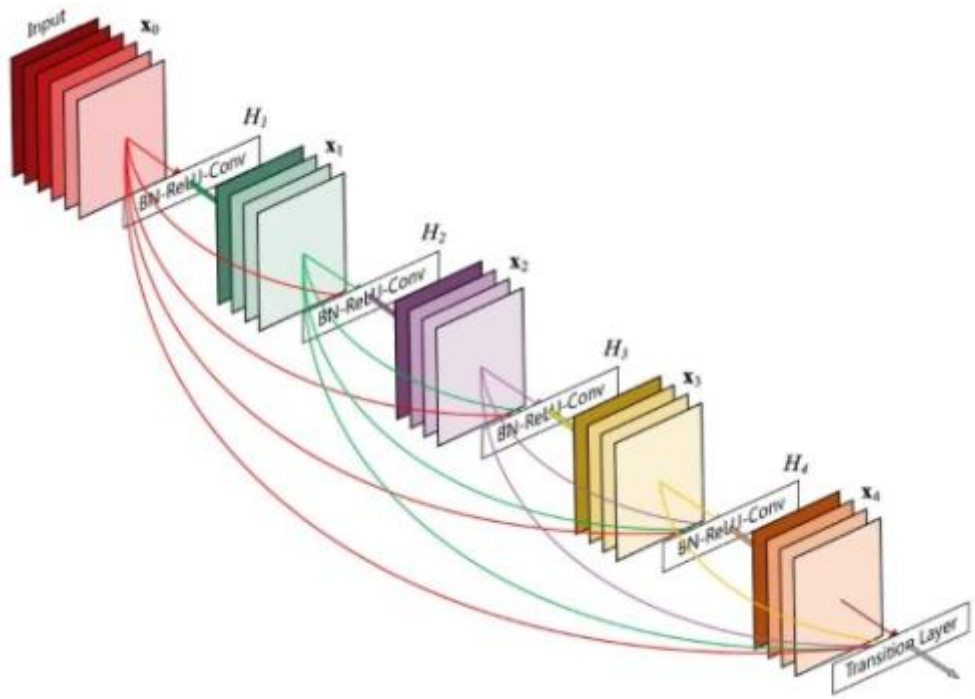


Hình 2 - 10: Hình ảnh bộ phát hiện đối tượng [20]

2.1.6.1. Dense Block & DenseNet

Để tăng độ chính xác của mô hình deep learning thì ta có thể tăng độ phức tạp của mô hình, mà cục thể là làm cho mô hình ngày càng sâu hơn. Tuy nhiên, việc này làm tăng độ phức tạp tính toán khi training, vì vậy, ta có thể áp dụng kỹ thuật như

skip-connection. Một kỹ thuật mở rộng của skip-connection là dense block. Dense block bao gồm nhiều lớp conv x(i) và H(i). Mỗi lớp H(i) bao gồm batch normalization, ReLU và theo sau bởi một conv layer. Các lớp H(i) này thay vì lấy input của layer ngay trước nó thì sẽ lấy tất cả các output của các layer trong dense block đó làm input hình ảnh được minh họa như hình 2-10.



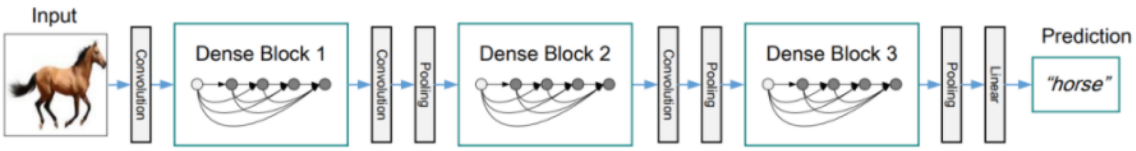
Hình 2 - 11: Kiến trúc mạng của Dense Block[6]

DenseNet có nhiều điểm giống với các mạng ResNet. Tuy nhiên, DenseNet có thể giải quyết được vấn đề vanishing gradient tốt hơn so với ResNet. DenseNet có một mẫu kết nối đơn giản để đảm bảo lưu lượng thông tin giữa các layer trong tính toán xuôi và tính toán gradient ngược là tối đa. Mạng này kết nối tất cả các layer lại, do đó mỗi layer đều được nhận thêm input từ các layer trước nó và truyền lại feature maps tới tất cả các layer sau nó. DenseNet-121 chỉ gồm 8 triệu tham số, nhưng nó đạt được độ chính xác cao hơn ResNet-50 với hơn 26 triệu tham số trên tập dữ liệu

[6] https://dothanhblog.wordpress.com/2020/05/08/yolov4/

ImageNet. DenseNet cũng có nhiều biến thể như DenseNet-121, DenseNet-169, DenseNet-201, DenseNet-264,…

Các khối dày đặc trong đó mỗi lớp được kết nối với mọi lớp khác theo kiểu chuyển tiếp. Làm giảm tình trạng vanishing gradient, tăng cường lan truyền đặc trưng cho những lớp phía sau, khuyến khích sử dụng lại các đặc trưng đã học. Tất cả những đặc điểm trên góp phần củng cố lượng kiến thức qua nhiều lần học mà không bị bão hòa theo độ sâu của mô hình. Qua đó cho thấy rằng chỉ với mạng nông chứa 50 lớp có thể hoạt động tốt hơn mạng sâu, điển hình là mạng ResNet-50 có hiệu quả vượt trội hơn hẳn ResNet-110.



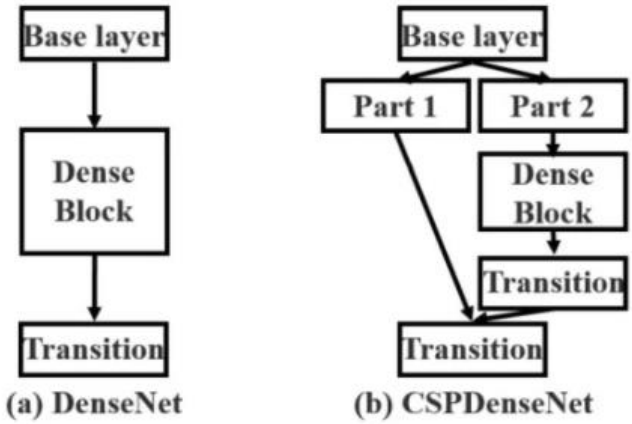
Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv, 3 × 3 conv] × 24	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 48	[1 × 1 conv, 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv, 3 × 3 conv] × 16	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Hình 2 - 12: Tổng quan về mạng sâu DenseNet[7]

[7] https://iq.opengenus.org/architecture-of-densenet121/

2.1.6.2. Cross-stage-partial-connection (CSP)

CSPNet chia input feature map ra thành hai phần bằng nhau, một phần giữ nguyên để đưa vào transition block, phần còn lại được đưa vào 1 dense block + 1 transition layer. CSP connection giúp cho vừa lưu giữ được một phần thông tin từ các layer trước, vừa giảm độ phức tạp của mô hình.



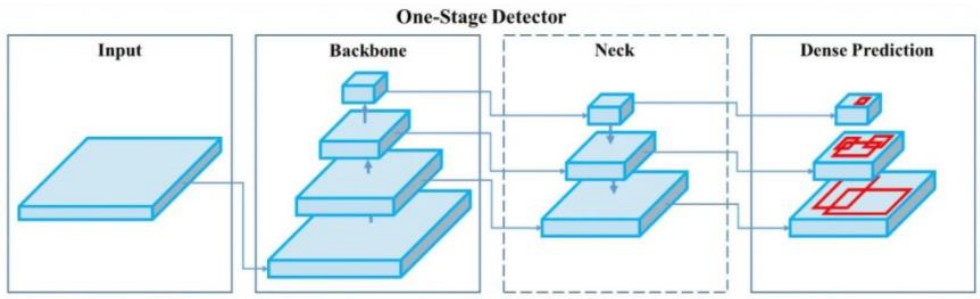
Hình 2 - 13: Sự khác nhau giữa DenseNet và CSPDenset[8]

2.1.6.3. SDPDarknet53

Yolov4 sử dụng CSPDarknet53 để làm backbone do CSPDarknet53 có độ chính xác trong task object detection cao hơn so với ResNet và mặc dù ResNet có độ chính xác trong task classification cao hơn nhưng hạn chế này có thể được cải thiện nhờ vào hàm activation Mish và một số kỹ thuật khác.

[8] https://jonathan-hui.medium.com/yolov4-c9901eaa8e61

2.1.6.4. Neck



Hình 2 - 14: Bộ phát hiện one-stage[9]

Một object detetector bao gồm một backbone (feature extraction) và một head (object detection). Để phát hiện đối tượng với các kích thước khác nhau, một kiến trúc mạng sử dụng các feature maps tại các vị trí khác nhau để predict.

Để làm giàu thông tin đẩy về head, một số feature map gần nhau trong bottom-up stream và top-down stream được kết hợp với nhau trước khi đẩy và head. Vì vậy, head sẽ giàu không gian thông tin (rich spatial information) từ bottom-up stream và giàu thông tin ngữ nghĩa (rich semantic information) từ top-down stream. Thành phần này được gọi là Neck.

Đặc trưng kim tự tháp đã trở thành một phần thiết yếu trong các mô hình phát hiện đối tượng hiện đại. Phương pháp thịnh hành nhất cho việc tạo ra các đặc trưng kim tự tháp là feature pyramid network (FPN). Các FPN có hai đặc điểm nổi bật:

- Tính năng kết hợp đa quy mô (Multi-Scale feature fusion): Kết hợp nhiều đặc trưng có độ phân giải cao (high-resolution) và đặc trưng có độ phân giải thấp (low-resolution) để tạo ra các biểu diễn đối tượng tốt hơn.
- Phân chia và chinh phục (Divide-and-conquer): Sử dụng các cấp độ đặc trưng khác nhau để phát hiện các đối tượng với các quy mô khác nhau.

Tác giả của You Only Look One-level Feature (YOLOF) đã nghiên cứu ảnh hưởng của hai lợi ích của FPN và thấy được chức năng phân chia và chinh phục của các FPN phần lớn đã bị bỏ qua. Do đó họ đã thiết kế các thử nghiệm tách rời sự kết hợp

[9] https://becominghuman.ai/explaining-yolov4-a-one-stage-detector-cdac0826cbd7

đặc trưng đa quy mô (multi-scale feature) và các chức năng của phân chia và chính phục (divide-and-conquer) với RetinaNet.

2.2. Các phương pháp two-stage

2.2.1. Giới thiệu phương pháp two-stage

Trong phương pháp two-stage các vùng đối tượng gần đúng được đề xuất bằng cách sử dụng các đặc trưng sâu trước khi các đặc trưng này được sử dụng để phân loại cũng như hồi quy bounding box cho đối tượng. Nói cách khác, đây là phương pháp xuất hiện giai đoạn khởi tạo vùng đề xuất (region proposal) để làm nền tảng cho quá trình huấn luyện thay vì tính toán thẳng trên feature map đầu vào nhằm hướng tới quá trình huấn luyện chất lượng hơn và ít sai sót nhờ vào sự giám sát của các vùng đề xuất.

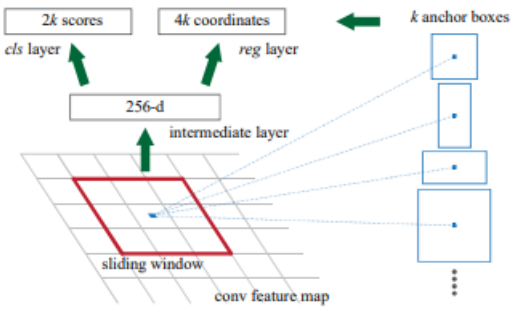
Kiến trúc two-stage đề xuất vùng đối tượng với các phương pháp Thị giác máy tính thông thường hoặc mạng sâu, sau đó phân loại đối tượng dựa trên các tính năng được trích xuất từ vùng được đề xuất với hồi quy bounding box. Các vùng đề xuất có thể được tạo ra nhờ những thuật toán với các thông số đã được thiết lập sẵn hoặc nâng cao hơn, chúng được tạo ra hoàn toàn nhờ vào một mạng Deep Learning được xây dựng để phục vụ cho nhiệm vụ này. Nhờ vậy, mô hình gốc có thể tận dụng lượng thông tin và đặc trưng bên trong những vùng đề xuất – những vùng có thể chứa đối tượng – để điều hướng sự tập trung đến đối tượng một cách hiệu quả và đưa ra những dự đoán chính xác.

Phương pháp two-stage đạt được độ chính xác phát hiện cao nhưng thường chậm hơn. Do có nhiều bước suy luận trên mỗi hình ảnh, hiệu suất (khung hình trên giây) không tốt bằng các phương pháp one-stage. Hơn nữa, kiến trúc two-stage đặc thù đòi hỏi phần tạo ra các vùng đề xuất rất nhiều thời gian để tạo ra các vùng chất lượng tốt, đồng thời còn phải ứng dụng tính toán tích chập với lượng tham số không nhỏ trên các vùng này. Ban đầu, với số lượng vùng đề xuất còn lớn, con người đã

cải tiến và ngày càng tối ưu quy trình này hơn qua việc bỏ đi các thông số mặc định trong thuật toán mà chuyển dần sang ứng dụng tiềm năng của mạng học sâu. Từ đó, tốc độ lẫn hiệu năng được cải thiện đáng kể, nhưng về mặt bằng chung, tốc độ vẫn không thể nhanh hơn các kiến trúc one-stage, nhờ vào kiến trúc gọn hơn trong quá trình xây dựng.

2.2.2. Faster R-CNN

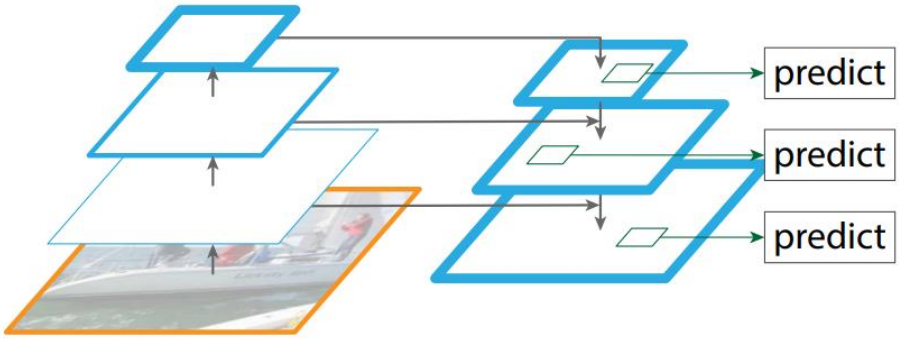
Faster R-CNN [14] là mô hình tốt nhất của họ nhà R-CNN, được thiết kế thêm 1 mạng con gọi là RPN (Region Proposal Network) thay cho Selective Search để trích rút các vùng có khả năng chứa đối tượng của ảnh, các phần sau đó được thực hiện tương tự như Fast-RCNN nhưng nhanh hơn nhiều và được thiết kế như 1 mạng end-to-end trainable network. Region Proposal Network nhận đầu vào là ảnh với kích thước bất kì và cho đầu ra là region proposal, cùng với xác suất vật thể có thể tồn tại ở vị trí đó. RPN được xây dựng với 2 công đoạn chính: tiến hành nạp ảnh vào Deep Neural Network(DNN) để thu được các đặc trưng tích chập (convolutional features) và sau đó sử dụng cơ chế cửa sổ trượt (sliding window) lên những đặc trưng mà ta đã thu được từ trước. Với cấu trúc như thế, Faster R-CNN đạt tốc độ nhanh hơn rất nhiều so với Fast R-CNN, chênh lệch khoảng 20 lần (2 giây mỗi hình với Faster R-CNN và 42 giây mỗi hình với Fast R-CNN).



Hình 2 - 15: Region Prosal Network

2.2.3. Feature Pyramid Networks

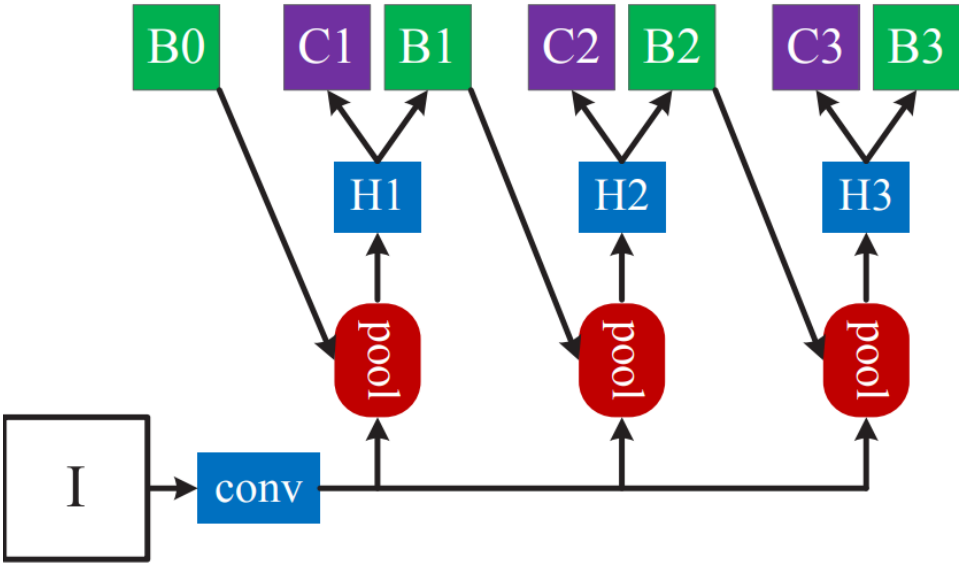
Trong lĩnh vực thị giác máy tính việc nhận diện các đối tượng ở các kích thước khác nhau là một thách thức lớn. Feature Pyramid Networks hay gọi tắt là FPN [7] được đề xuất bởi Lin et al với một kiến trúc top-down kết hợp với các kết nối cạnh bên (lateral connections), mạng đã tận dụng tối đa được các đặc trưng ngữ nghĩa cấp cao (high-level semantic feature map) ở mọi kích thước. Ở đường bottom-up độ phân giải sẽ giảm, nhưng giá trị ngữ nghĩa sẽ tăng lên. Ngược lại đường top-down nhằm mục đích xây dựng các layer có độ phân giải cao từ các layer có độ phân giải cao từ các player có ngữ nghĩa cao. Nhờ vậy, FPN đã cho thấy những sự cải thiện đáng chú ý trong nghiên cứu lẫn ứng dụng của pyramid features.



Hình 2 - 16: Cấu trúc của Feature Pyramid Networks

2.2.4. Cascade R-CNN

Cai và Vasconcelos [2] đã đề xuất Cascade R-CNN là bộ phát hiện đối tượng chất lượng cao, các head khác nhau được sử dụng ở các tầng khác nhau. Mỗi head được thiết kế cho một ngưỡng IoU cụ thể từ nhỏ đến lớn để tránh false positive. Cascade R-CNN giảm được tình trạng overfitting trong quá trình huấn luyện và giảm sự không khớp về chất lượng trong thời gian suy luận. Hơn nữa, nhờ vào các ngưỡng IoU tăng dần trong kiến trúc, mô hình hoàn toàn có thể lọc được các bounding box gây nhiễu hoặc có độ chính xác không cao, từ đó có thể cải thiện hiệu suất hoạt động và kết quả sau cùng.

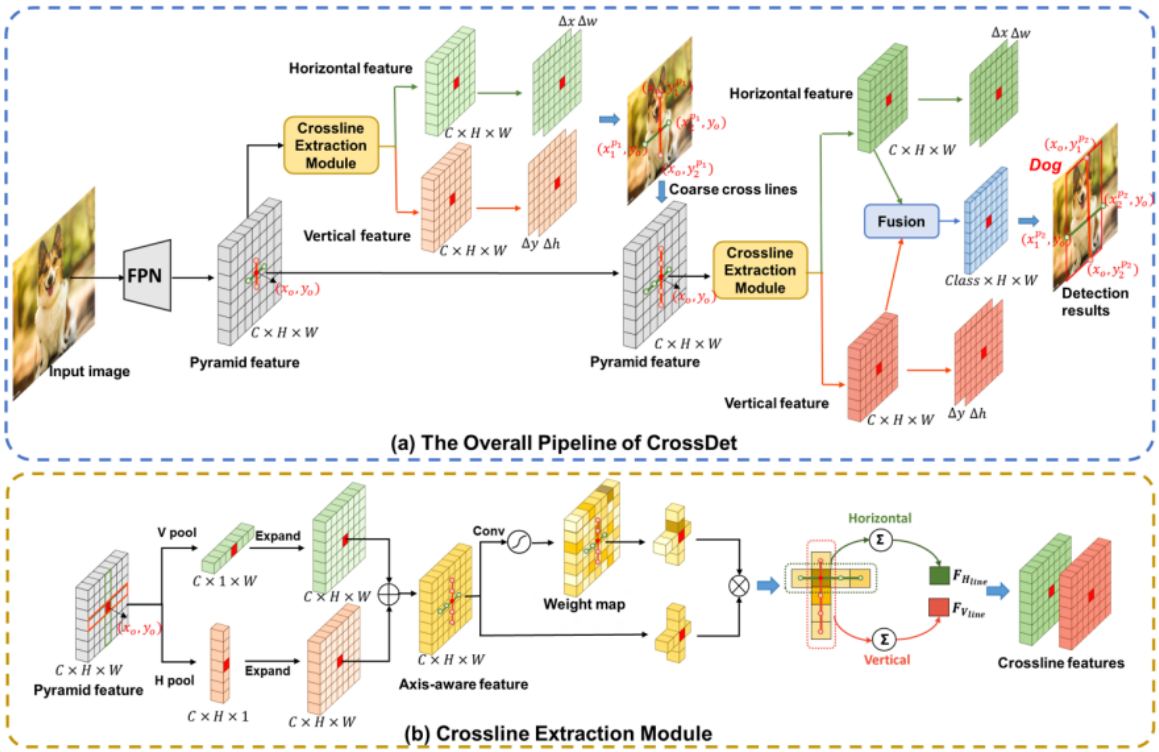


Hình 2 - 17: Cascade R-CNN

2.2.5. CrossDet

Với mục tiêu của **CrossDet** [13] là đảm bảo độ recall tốt, các phương pháp sử dụng anchor base phải thiết kế kiến trúc với rất nhiều siêu tham số khác nhau. Hơn nữa, các phương pháp này còn trích xuất toàn bộ đặc trưng bên trong bounding box điều này dẫn đến khi có hai vật thể chồng lên nhau, chúng dễ gây nhầm lẫn vì hai bounding box có phần giao nhau, gây ra những hạn chế nhất định cho tốc độ và khả năng phát hiện chính xác của mô hình. Trong khi đó, các phương pháp point-based sử dụng các điểm rời rạc để biểu diễn vật thể đang quan tâm, hạn chế được những điểm bất lợi của các phương pháp anchor-based. Tuy nhiên, những điểm rời rạc và riêng biệt này lại gây ra một tình trạng là mất mát lượng thông tin tiếp diễn, liên tục giữa các đặc trưng trong ảnh trong quá trình huấn luyện. Nhận thấy những khuyết điểm này của các phương pháp trước nhóm tác giả đã giới thiệu CrossDet, một ancor-free detector, sử dụng một bộ các đường chữ thập (cross line) để biểu diễn vật thể, giúp phương pháp này xác định vật thể linh động và hiệu quả hơn. Vì không có các siêu tham số như anchor box, các cross line được phát triển theo chiều ngang và chiều dọc với sự giám sát của vị trí, chiều rộng và chiều cao của ground-truth. Kiến

trúc của CrossDet được xây dựng với một multi-stage pipeline như hình 2-20 sử dụng backbone FPN (feature pyramid network) để tận dụng được tiềm năng của đặc trưng đa cấp độ. CrossDet gồm 2 giai đoạn là tạo các đường sơ khai và hồi quy các đường thẳng đó và bọc lại bằng bounding box. Ở stage đầu tiên trước khi có bộ cross line hoàn thiện, các cross line sơ khai được tạo ra với độ dài là 3 pixel sau đó mở rộng dần ra. Ở stage tiếp theo, để trích xuất các đặt trưng quan trọng, crossline extraction module (CEM) gồm hai phần axis-aware pooling và cross-line sampling. Axis-aware pooling thực hiện với band window (1, W) hoặc (H, 1) để mã hóa thông tin và tính năng theo trục ngang và trục dọc sau đó, mở rộng kích thước của feature map dạng dải nhận được thành $C \times H \times W$ và kết hợp chúng lại để tăng cường các đặc trưng gốc. Crossline sampling dựa vào feature map I' đã được tích hợp trước đó, tiến hành khởi tạo weight map tương ứng bằng 1×1 conv và một normalization layer sử dụng hàm sigmoid sau đó tiến hành lấy mẫu đặc trưng crosslines (horizontal, vertical).

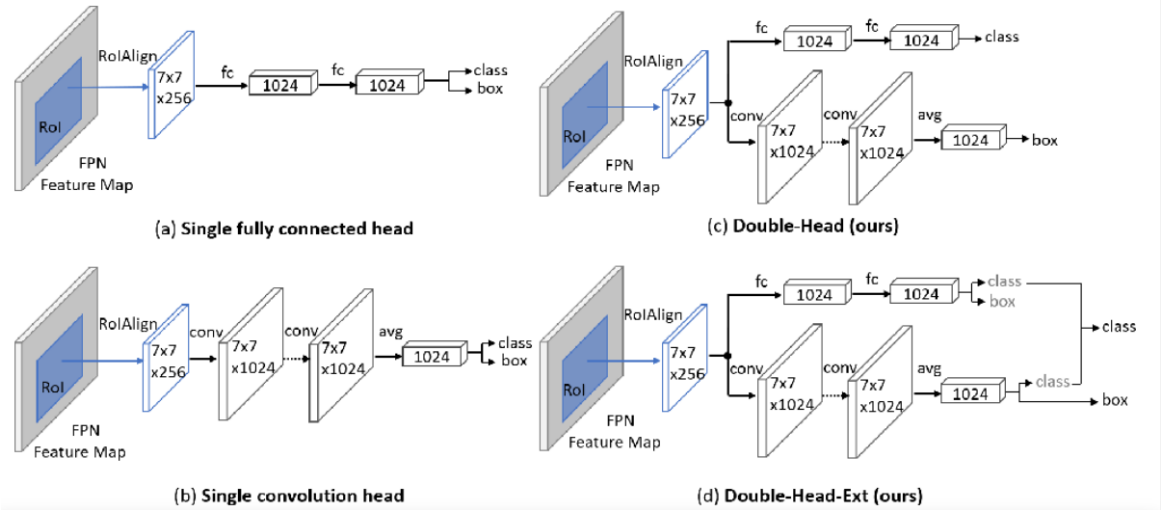


Hình 2 - 18: Cấu trúc của CrossDet.

Chương 3. **Phương pháp đề xuất cải thiện hiệu suất của RCNN**

3.1. Double Head RCNN

Double-Head RCNN [21] được xây dựng dựa trên kiến trúc Feature Pyramid Network (FPN) như tại hình 3-1. Trái ngược với các phương pháp đi trước, những phương pháp mà chỉ sử dụng một head duy nhất để rút trích đặc trưng trong vùng quan tâm RoI cho cả bài toán phân lớp và hồi quy bounding box, Double-head RCNN có kiến trúc chia thành 2 head riêng biệt tương thích cho từng bài toán classification và localization.



Hình 3 - 1: Sự khác nhau giữa single-head và double-head.

Qua quá trình nghiên cứu lẫn thực nghiệm, tác giả đã thấy được sự chênh lệch classification score ở các ngưỡng IoU của fully-connected head (fc-head) một cách rõ rệt và hơn hẳn so với convolution head (conv-head). Tác giả nhận định rằng fc-head có những sự nhạy cảm nhất định về không gian (spatial sensitivity) và nhờ đó, mô hình có thể dễ dàng phân biệt các bộ phận hoặc những thành phần giữa các vật thể khác nhau nhưng fc-head không quá nổi trội trong việc xác định miền offset của toàn vật thể. Ngược lại, các shared transformation (convolutional kernels) hoạt động rất hiểu quả trong conv-head trên mọi vị trí của feature map đầu vào, và sau đó sử dụng average pooling để tổng hợp.

Thông qua nghiên cứu của tác giả, fc-head xử lý bài toán classification tốt hơn conv-head và ngược lại conv-head xử lý bài toán bounding box regression tốt hơn fc-head. Từ đó, việc thiết kế ra kiến trúc double-head (một sự kết hợp giữa conv-head và fc-head) sẽ tạo tiền đề để mang lại kết quả nổi bật trong quá trình giải quyết bài toán classification và localization.

Lý do mà fc-head không hoàn toàn thích hợp với bài toán localization vì đầu ra của fc-head không phải là một feature map (thứ sẽ mang đặc trưng ở nhiều level cũng như là thông tin liên tục của đối tượng) mà là một vector đặc trưng, từ đó nhận thấy rằng fc-head hồi quy bounding box không quá tốt vì lượng thông tin liên tục bị thiếu hụt. Nhờ vào vector đặc trưng của fc-head, bài toán classification được thực hiện một cách quyết đoán hơn và ổn định hơn với những đặc trưng ở cấp độ cần thiết.

Do đó, kiến trúc double-head sẽ tận dụng tối đa hiệu quả của fc-head để giải quyết bài toán classification và conv-head để giải quyết bài toán localization. Đối với kiến trúc Double-head-ext (một phiên bản nâng cấp của Double-Head R-CNN), fc-head và conv-head sẽ hỗ trợ lẫn nhau trong bài toán classification vì tác giả tin rằng tuy kiến trúc của fc-head và con-head khác nhau nhưng lượng thông tin thu được có thể bổ trợ lẫn nhau bởi kết quả nghiên cứu của họ chỉ ra rằng cả hai head đều đạt được sự hiệu quả nhất định cho bài toán classification (chỉ là fc-head có chút vượt trội hơn conv-head nhưng sự chênh lệch đó là không đáng kể).

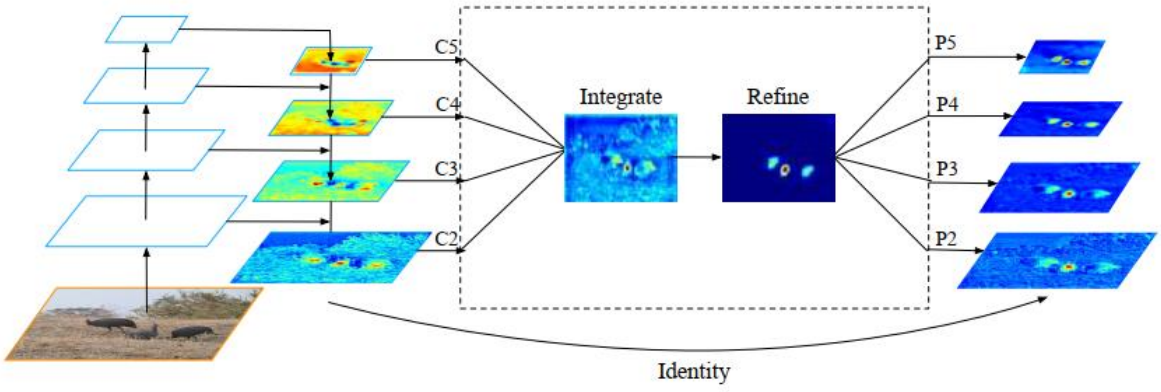
3.2. Libra RCNN

Pang et al. [11] đề xuất Libra R-CNN dựa trên FPN Faster R-CNN với mục tiêu là hướng tới sự cân bằng trong quá trình huấn luyện. Libra RCNN tỏ ra hiệu quả vượt trội qua việc có thể vận hành và cải thiện kết quả trên hàng loạt backbone cho cả one-stage detector và two-stage detector. Có ba vấn đề mất cân bằng được rút ra từ các phương pháp trước đó: Sample level imbalance, feature level imbalance và objective level imbalance.

Với Libra RCNN, nó đã giải quyết được sự mất cân bằng một cách hiệu quả trong sample level imbalance, hơn nữa mô hình cũng ít tốn kém hơn các phương pháp đi

trước, cũng như các chi phí liên quan. Bên cạnh đó, đối với feature level imbalance, Libra RCNN đã rất hiệu quả trong việc tích hợp các đặc trưng ngữ nghĩa cân bằng (balanced semantic feature) nhằm thúc đẩy tăng cường các đặc trưng gốc, rút trích và tạo ra đặc trưng dễ nhận biết hơn của vật thể. Cuối cùng, Libra RCNN đã có những sự cải tiến, tinh chỉnh để khiến mạng này có thể đạt được sự hội tụ tốt hơn dành cho vấn đề objective level imbalance.

Kiến trúc Libra RCNN đã giải quyết lần lượt những sự mất cân bằng ở ba cấp độ nói trên. Ở sample level, tác giả đã thiết kế ra một hàm xác suất để có thể chọn ra một số lượng mẫu negative nhất định từ tổng mẫu ứng cử viên tương ứng. Nhận thức được rằng các hard sample sẽ có ích rất nhiều cho quá trình huấn luyện mô hình, tác giả đã tìm cách tăng xác suất chọn được các mẫu hard negative bằng một hàm xác suất mới sau khi phân những mẫu đã chọn ở bước trước vào ba ngăn (mặc định của Libra RCNN) giúp tăng khả năng chọn lọc các đặc trưng tốt, chất lượng cho quá trình huấn luyện. Để sự cân bằng trong sample level được toàn diện hơn, mạng sẽ lấy số lượng mẫu positive bằng nhau trong mỗi ground-truth, thứ mà bọc lấy vật thể. Theo sau sample level là feature level, sự mất cân bằng ở đây diễn ra là do lượng đặc trưng thu được ở các từng tầng (điển hình là trong kiến trúc của FPN) không đồng đều dẫn tới sự chênh lệch thông tin và đặc trưng. Do đó, tác giả đã tiến hành cân bằng lượng thông tin tương ứng với từng tầng kiến pyramid để thu được balanced semantic feature thông qua các phép biến đổi nhất định như hình 3 - 2.



Hình 3 - 2: Hình ảnh pipeline và heatmap of balanced feature pyramid.

Ngoài ra, để điều chỉnh các feature thu được, tác giả đã sử dụng hai phương thức là thực hiện tích chập trực tiếp hoặc sử dụng một non-local module. Sau cùng, vì các non-local module đạt được sự ổn định nhất định nên tác giả đã quyết định sử dụng nó trong phương pháp này. Với phương pháp trên, các đặc trưng ở low-level và cả high-level đều được tổng hợp cùng lúc và đảm bảo được sự cân bằng ở feature level. Cuối cùng là trong objective level, sự mất cân bằng gây ra bởi sự chênh lệch lượng gradient đóng góp từ các sample, bao gồm easy sample và hard sample. Theo nghiên cứu, các hard samples sẽ tạo ra một lượng gradient quá tải và gây hại cho quá trình huấn luyện. Trong khi đó, các easy sample lại đóng góp ít gradient hơn đến tổng lượng gradient so với những hard samples. Do đó, mục tiêu mà tác giả hướng tới ở đây là cân bằng các sample cho từng bài toán nhất định, cụ thể hơn là đẩy mạnh các gradient được tạo ra từ cái easy sample. Tác giả đã đề xuất hàm balanced L1 Loss để cân bằng tác động của hai task, là classification và localization, và cả lượng gradient đóng góp của các sample thu được. Nói cách khác, mức độ quan trọng cũng như là tầm ảnh hưởng hai bài toán bây giờ là như nhau, các sample sẽ được ‘đối xử’ công bằng để tận dụng tối đa lượng thông tin mà một bức ảnh đầu vào có thể mang lại.

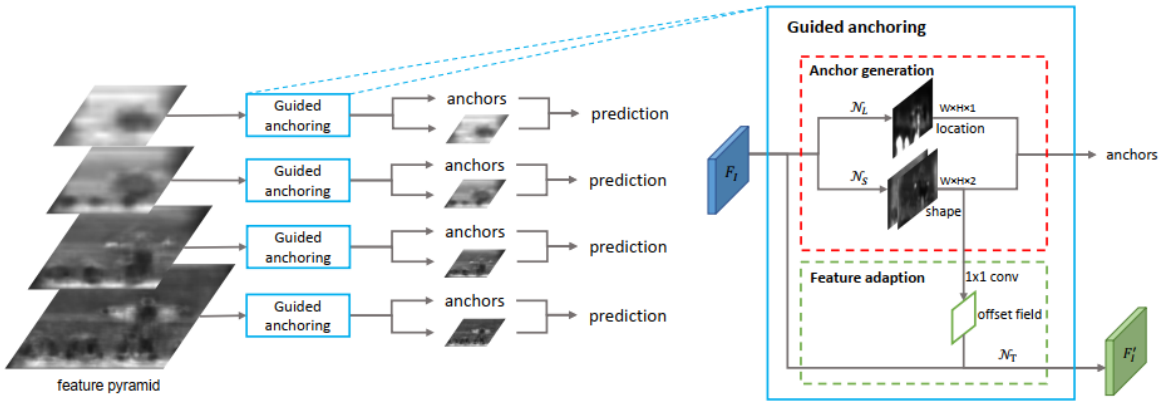
3.3. Guided Anchoring

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.



Hình 3 - 3: Framework của cơ bản Guided Anchoring.

Đối với nhánh Anchor Shape Prediction, với tọa độ (x, y) 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 (w, h) 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 1×1 convolution để tạo ra một ánh xạ phù hợp.

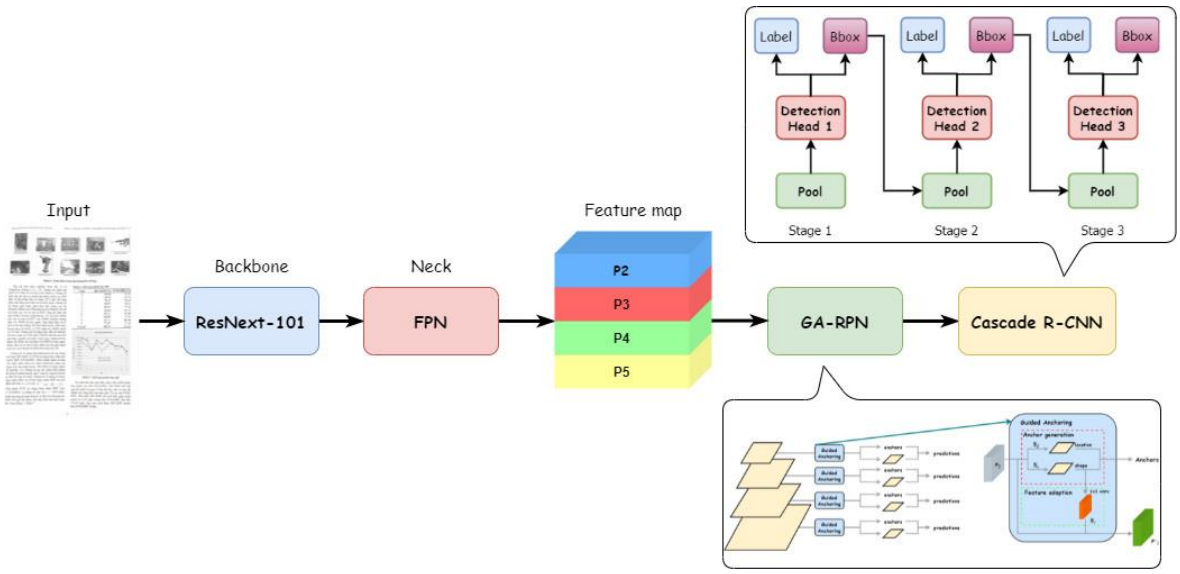
Cần phải nhấn mạnh rằng, thiết kế này hoàn toàn khác xa với các cơ chế khởi tạo anchor thông thường vì trong Guided Anchoring, mỗi vị trí chỉ liên kết với một anchor được dự đoán kèm theo kích thước đa dạng thay vì liên kết với một bộ các anchor có kích thước được định nghĩa sẵn. Để khai thác và trích 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.

3.4. Phương pháp đề xuất Guided Anchoring Cascade R-CNN

Như đã đề cập tại ở các phần trước đó, Double Head R-CNN, Libra R-CNN hay Guided Anchoring là những thiết kế có thể được xem như là ‘phụ kiện’ và có thể được bổ sung vào để cải thiện kết quả của mô hình Faster R-CNN. Do đó, những

'phụ kiện' này có thể được xem một mô-đun có thể dễ dàng gán cho các máy dò khác, cụ thể ở đây các tác giả đã sử dụng Faster R-CNN. Bên cạnh đó, Cascade R-CNN là một phần mở rộng nhiều giai đoạn của R-CNN trong đó các giai đoạn của máy do được chọn lọc tuần tự hơn qua đó đã cải thiện được những hạn chế của Faster R-CNN như đã được đề cập tại ở trước đó. Vì thế, chúng tôi đã đề xuất mô hình một hình dựa trên kiến trúc Cascade R-CNN kết hợp với Guided Anchoring để nâng cao hiệu suất phát hiện đối tượng trên tài liệu dạng ảnh.



Hình 4 - 1: Kiến trúc Guided Anchoring Cascade R-CNN

Nhận thấy được tiềm năng của các đặc trưng ở nhiều cấp độ (multi-level feature), chúng tôi quyết định sẽ tận dụng FPN trong phương pháp này, từ đó thu được lượng thông tin quý giá được thể hiện dưới các bộ feature map với các kích thước tương ứng với các tầng đã được xây dựng trong kiến trúc từ trước. Nhưng trước đó, chúng tôi lấy ResNeXt-101 làm backbone cho mô hình, kiến trúc được mô tả như trên hình 4-1. Không chỉ lượng đặc trưng dồi dào thu được từ FPN, chúng tôi còn tiến hành áp dụng Guided Anchoring để điều hướng và khởi tạo được các anchor box chất lượng và linh hoạt về kích thước, nhằm tăng khả năng phát hiện được các đối tượng

có kích thước khá khác biệt. Đồng thời, phương pháp này còn giúp chúng tôi giảm được lượng anchor box tạo ra so với các phương pháp trước, qua đó giảm được những chi phí và tăng tốc độ xử lý trong quá trình huấn luyện một cách đáng kể.

Cuối cùng, khi đã có đủ lượng nguyên liệu cần thiết cho quá trình huấn luyện của bài toán DOD, chúng tôi sẽ nạp những nguyên liệu này vào mạng Cascade R-CNN. Nhờ vào các anchor chất lượng tốt đến từ quá trình khởi tạo của Guided Anchoring kết hợp với khả năng lọc bounding box không cần thiết từ việc tác động lên ngưỡng IoU lên từng tầng trong kiến trúc của Cascade R-CNN, chúng tôi tin rằng kết quả sẽ được cải thiện đáng kể. Bộ dữ liệu mà chúng tôi sẽ khảo sát và nghiên cứu sẽ là bộ dữ liệu tài liệu dạng ảnh tiếng Việt UIT-DODV nhằm phục vụ cho quá trình thực nghiệm và đánh giá kết quả mô hình. Trong Cascade R-CNN, với nguyên lý hoạt động là tăng dần ngưỡng IoU theo từng giai đoạn của quá trình huấn luyện, do đó có sự xuất hiện của các Detection head (được đánh số là 1, 2, 3) được thiết kế tương ứng với từng tầng, cụ thể hơn là tương ứng với từng ngưỡng IoU.

Chương 4. **Thực nghiệm và đánh giá**

4.1. Bộ dữ liệu UIT-DODV

UIT-DODV là bộ dữ liệu ảnh tài liệu đầu tiên của Việt Nam, bao gồm 2.394 ảnh với 4 lớp: Bảng, Hình, Chú thích, Công thức. UIT-DODV đã chuyển đổi 1.696 hình ảnh từ PDF với kích thước 1.654×2.338, 247 hình ảnh được scan từ máy scan và 451 hình ảnh được scan từ điện thoại thông minh.

UIT-DODV có những điểm nổi bật sau:

- Hình ảnh đa dạng: hình ảnh trong bộ dữ liệu có hai loại, với hình ảnh được chuyển đổi từ PDF dưới dạng tài liệu và hình ảnh. Ảnh scan thường có độ phân giải thấp hơn tùy thuộc vào góc scan cũng như điều kiện ánh sáng có thể khiến trang tài liệu bị nhòe, méo, lệch, che khuất.
- Bố cục đa dạng: dữ liệu được thu thập từ các hội nghị / tạp chí khoa học khác, đặc điểm chung của các hội nghị tạp chí này là thường sử dụng các

mẫu của chúng (thông thường các trang tài liệu có thể thể hiện các trang tài liệu dưới dạng một cột hoặc hai cột).

- Thách thức đến từ các lớp dữ liệu: với việc sử dụng đồng thời hai đối tượng công thức (Formula) và Chú thích cũng tạo ra một thách thức cho tập dữ liệu. Như trong việc xây dựng mô hình phát hiện cho các đối tượng này. Phần lớn trang tài liệu được biểu diễn dưới dạng văn bản, vì vậy các đối tượng khó được phát hiện nhanh chóng, thậm chí còn bị phân loại nhầm hoặc bỏ sót, ảnh hưởng trực tiếp tới kết quả và hiệu suất của các phương pháp.

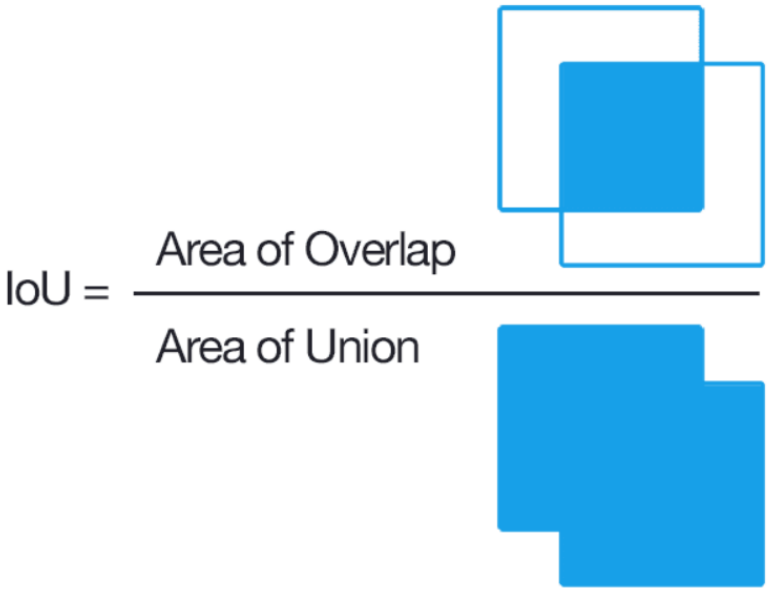
4.2. Các độ đo đánh giá

4.2.1. Intersection over Union (IoU)

Intersection over Union là chỉ số đánh giá được sử dụng để đo độ chính xác của phát hiện đối tượng trên tập dữ liệu cụ thể. Cụ thể hơn, đây là chỉ số thể hiện sự trùng lặp của bounding box được dự đoán so với ground-truth bounding box. IoU nhận giá trị cao nhất là 1, tức là mô hình dự đoán đúng tuyệt đối, về mặt cơ bản, khi IoU càng cao, mô hình càng tốt. Chỉ số này thường được gặp trong các Object Detection Challenge và trong các bài toán thực tế. IoU thường được sử dụng để đánh giá hiệu năng của các bộ phát hiện đối tượng như HOG + Linear SVM và mạng nơ ron tích chập (R-CNN, Fast R-CNN, YOLO,…).

Để áp dụng được IoU để đánh giá cần:

- Đường bao thực (ground-truth bounding box): là đường bao mà chúng ta gán cho vật thể bằng labelImg tool.
- Đường bao dự đoán (predicted bouding box): là đường bao chúng ta sử dụng file Weights sau khi đào tạo để nhận dạng.



Hình 4 - 2: Minh họa IoU[10]

4.2.2. Average Precision

The average precision (AP) là cách để thể hiện đường cong precision-recall thành một giá trị duy nhất đại diện cho giá trị trung bình của tất cả precisions. Độ đo AP được tính theo công thức dưới đây. Sử dụng một vòng lặp đi qua tất cả precisions/recalls, Sự khác biệt giữa recall hiện tại và recall tiếp theo tính được, sau đó nhân với precision hiện tại. Nói cách khác, AP là tổng precision ở mỗi threshold trong đó trọng số được tăng trong recall.

$$AP = \sum_{k=0}^{k=n-1} [Recalls(k) - Recalls(k+1) * Precisions(k)$$

$$Recalls(n) = 0, Precisions(n) = 1$$

$$n = số\ lượng\ threshold$$

[10] https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

Đối với mỗi lớp đối tượng $c_i \in C = \{1, \dots, n\}$, ta xây dựng một confusion matrix sao cho confusion matrix thứ i lấy lớp ci làm lớp positive và tất các lớp c_j với $j \neq i$ là các lớp negative. Các công thức bên dưới kí hiệu TP_i, TN_i, FP_i, FN_i lần lượt là số các điểm dữ liệu được dự đoán là True Positive, True Negative, False Positive và False Negative trong confusion matrix tương ứng với lớp đối tượng thứ i.

Confusion matrix là một phương pháp đánh giá kết quả những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp. Một confusion matrix gồm 4 chỉ số sau đối với mỗi lớp phân loại:

True Positive: Số lượng dự đoán chính xác.

False Positive: Số lượng dự đoán chính xác một cách gián tiếp.

True Negative: Số lượng các dự đoán sai lệch.

False Negative: Số lượng các dự đoán sai lệch một cách gián tiếp.

Tuy nhiên, trong bài toán phát hiện đối tượng, đơn vị True Negative sẽ không được xem xét đến.

4.2.2.1. Precision

Precision được định nghĩa là tỉ lệ số sample được tính là True Positive(TP) với tổng số sample được phân loại là Positive (TP + FP). Precision càng lớn có nghĩa là độ chính xác sẽ càng cao. Precision có giá trị trong khoảng (0,1).

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

4.2.2.2. Recall

Recall được định nghĩa là tỷ lệ giữa các điểm positive thực được nhận đúng trên tổng điểm positive thực (TP + FN). Recall cao nghĩa là tỉ lệ bỏ sót các sample positive thực thấp.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

4.2.2.3. Mean Average Precision (mAP)

Nhằm mục đích so sánh với kết quả baseline trên bộ dữ liệu UIT-DODV, chúng tôi sử dụng độ đo mP (mean Average Precistion) [8] để đánh giá. mAP là trung bình của AP trên tất cả các lớp đối tượng. Đối với mỗi lớp, tính AP ở các ngưỡng IoU khác nhau và lấy điểm trung bình của chúng để lấy AP của lớp đó. Sau đó sẽ tính mAP bằng cách lấy trung bình AP qua các lớp khác nhau. Bên cạnh đó, chúng tôi cũng sẽ báo cáo kết quả trên các độ đo AP_{50} và AP_{75} tương ứng với IoU threshold là 0.5 và 0.75.

Công thức tính mAP:$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$

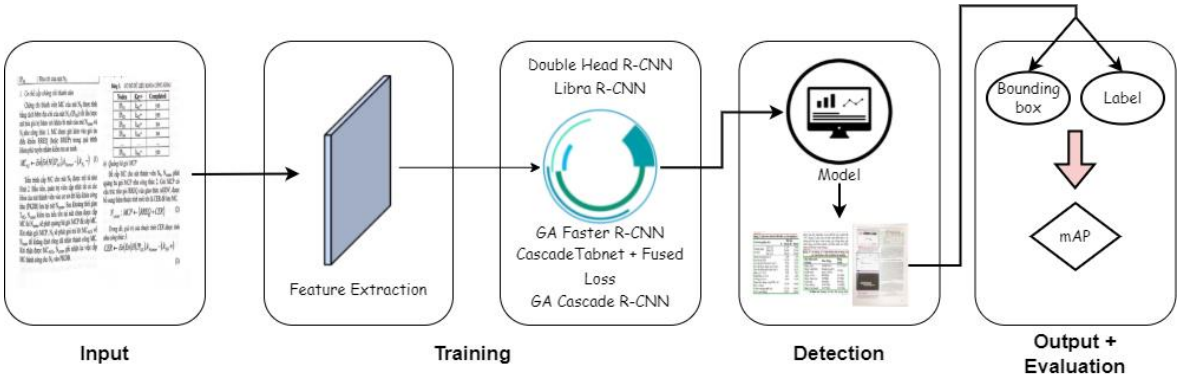
Trong đó:

$$AP_k: AP\ của\ lớp\ k$$

$$n: số\ lớp$$

4.3. Mô tả thực nghiệm

Bài toán phát hiện đối tượng trên tài liệu dạng ảnh được chúng tôi thực nghiệm theo quy trình quy trình như sau:



Hình 4 - 3: Quy trình thực nghiệm

Ban đầu, chúng tôi sẽ cho vào feature extraction để lấy các thông tin cần thiết của input đầu vào. Sau đó sẽ đưa vào các mô hình để huấn luyện. Với mô hình Double-Head R-CNN, Libra R-CNN, GA Faster R-CNN, CascadeTabNet + Fused Loss, Guided Anchoring + Cascade R-CNN, mô hình sẽ đánh giá trên tập validation sau mỗi epchs. Bộ trọng số tại epoch cho kết quả cao nhất sao khi kết thúc quá trình huấn luyện sẽ được dùng để làm trọng số dự đoán cho tập test.

Sau đó sử dụng các độ đo AP_{50}, AP_{75} và mAP để đánh giá độ hiệu quả của mô hình. Đầu ra cho mô hình là ảnh có chưa bouding-box (vị trí các lớp đối tượng), nhãn và độ tin cậy (confidence) của mô hình.

Cấu hình thực nghiệm

Chúng tôi triển khai thực nghiệm trên 2x GPU RTX 2080 Ti dựa trên framework MMDetection.

4.4. Kết quả thực nghiệm và đánh giá

4.4.1. Kết quả thực nghiệm

Chúng tôi tiến hành thực nghiệm 3 phương pháp (Double-Head R-CNN, Libra R-CNN, Guided Anchoring Faster R-CNN) và trích dẫn kết quả phương pháp CascadeTabNet + Fused Loss [5] để đối chiếu với hiệu năng của phương pháp mà

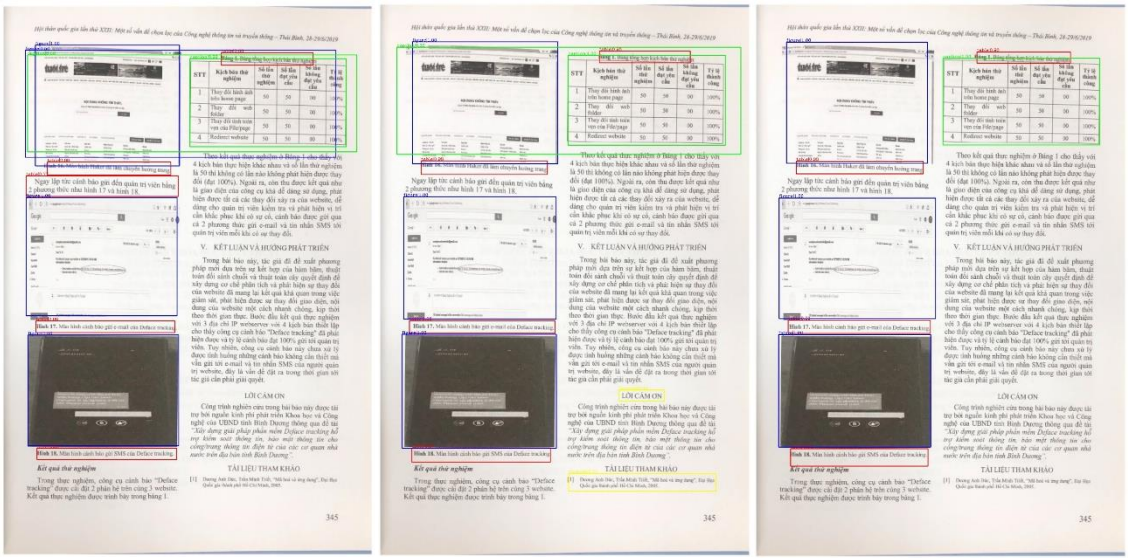
chúng tôi đề xuất. Kết quả được trình bày ở bảng 1.1 cho thấy kết quả rất khả quan của mô hình chúng tôi.

Bảng 1. 1 Kết quả thực nghiệm trên bộ dữ liệu UIT-DODV

Method	**Table**	**Figure**	**Caption**	**Formula**	**AP@.5**	**AP@.75**	**mAP**
Double-Head R-CNN	91.5	80.6	65.6	46.3	88.7	78.4	71.0
Libra R-CNN	92.5	81.0	68.2	46.0	89.6	79.1	71.9
Guided Anchoring Faster R-CNN	92.7	81.6	73.3	46.9	91.0	80.8	73.6
CascadeTabNet + Fused Loss [5]	94.3	83.0	73.7	47.5	89.1	81.6	74.5
Guided Anchoring Cascade R-CNN(Ours)	**95.4**	**84.8**	**75.9**	**50.5**	**91.8**	**83.1**	**76.6**

4.4.2. Trực quan hóa kết quả

Dưới đây là hình trực quan hóa kết quả dự đoán của ba phương pháp Doudle-Head R-CNN, Libra R-CNN, Guided Anchoring trên bộ dữ liệu UIT-DODV:

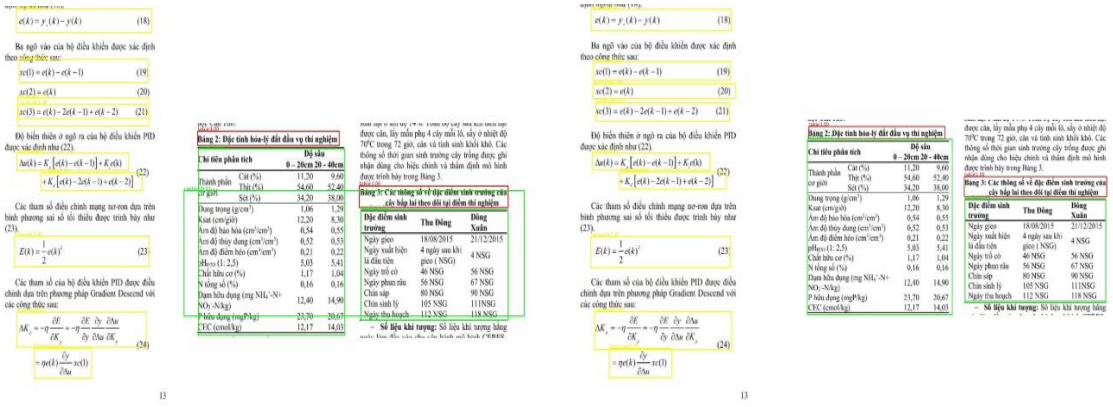


Double-Head R-CNN Libra R-CNN Guided Anchoring

Hình 4 - 4: Trực quan hóa kết quả dự đoán của 3 mô hình Double-Head-RCNN, Libra R-CNN và Guided Anchoring trên bộ dữ liệu UIT-DODV - (bbox màu xanh lá cây - bảng, màu xanh dương - hình, màu đỏ - chú thích và màu vàng - công thức)

Double-Head R-CNN phân loại các đối tượng khá chính xác khi mà 3 đối tượng như ảnh minh họa điều được phân loại đúng tuy nhiên vẫn còn hiện tượng overlap xảy ra ở hai lớp hình và bảng khi hai đối tượng này ở gần nhau. Cũng giống như Double-Head R-CNN, Libra R-CNN cũng xảy ra hiện tượng overlap ở bảng khi mà nó được đặt gần với hình, thêm vào đó Libra R-CNN phân loại nhầm văn bản bình thường thành công thức, những sai sót tuy nhỏ trên lại ảnh hưởng không hề ít tới kết quả chính xác cuối cùng của phương pháp. Tuy Guided Anchoring chưa thật sự hoàn hảo khi các bounding box còn chưa khớp hoàn toàn với các đối tượng nhưng sự sai sót không đáng kể, quan trọng là các đối tượng đã được dự đoán hoàn toàn chính xác mà không có overlap hay nhầm lẫn. Nhìn chung, Double Head R-CNN và Libra R-CNN đã có những kết quả rất tốt nhưng vẫn chưa tập trung được vào quá trình khởi tạo các anchor chất lượng cao dẫn tới độ hiệu quả bị giảm đáng kể. Với cơ chế điều hướng anchor với kích thước các anchor box linh hoạt trong Guided Anchoring, phương pháp đã rất thành công trong quá trình hình thành những hộp neo chất lượng, không quá dày đặc nhưng cũng không quá thưa thớt, loại bỏ hoàn toàn đi thông số cố định làm cho các vật thể, đối tượng có kích cỡ, hình dạng khác biệt vẫn có thể được bọc bởi bounding box dự đoán. Từ đó, Guided Anchoring có độ hiệu quả tốt nhất trong những phương pháp trên bởi nó đã tập trung được vào quy trình nền tảng của các phương pháp anchor-based, cụ thể hơn là tăng cường chất lượng của các anchor box mà không gây ra sự hao tốn nào về tài nguyên, chi phí so với các bộ phát hiện đối tượng trước.

Tiếp theo chúng ta có kết quả của Faster R-CNN và Cascade R-CNN sử dụng Double-Head R-CNN:



Faster R-CNN Cascade R-CNN

Hình 4 - 5: So sánh giữa Faster R-CNN và Cascade R-CNN sử dụng Double-Head-RCNN (bbox màu xanh lá cây - bảng, màu xanh dương - hình, màu đỏ - chú thích và màu vàng - công thức

Faster R-CNN phát hiện được hiểu quả của mình khi kết quả nhận được là những dự đoán rất chính xác các đối tượng. Tuy nhiên, tình trạng overlap xuất hiện khi hai bảng được đặt gần nhau, tương tự với đối tượng "Công thức". Hơn nữa, còn xuất hiện bỏ sót công thức như trên hình 4 - 3. Trong khi đó Cascade R-CNN không những phát hiện đúng các đối tượng mà các bounding box đầu ra tương đối khớp với các đối tượng. Theo chúng tôi, kiến trúc phân tầng của phương pháp Cascade R-CNN đã giải quyết rất hiệu quả vấn đề xảy ra khi tăng ngưỡng IoU, đó là hiệu suất của mô hình bị tác động làm giảm trình trạng bỏ sót, đồng thời phương pháp cũng tránh được khả năng overfitting.

4.4.3. Phân tích kết quả

Chúng tôi tiến hành đánh giá độ hiệu quả của bộ dữ liệu UIT-DODV trên 3 phương pháp là Double-Head R-CNN, Libra R-CNN, Guided Anchoring theo độ đo AP_{50}, AP_{75} và mAP. Kết quả thực nghiệm được chúng tôi tóm tắt trong bảng 1.1. Nhìn chung, kết quả tốt nhất thuộc về phương pháp common object detection cho thấy

Guided Anchoring với độ đo AP_{50}, AP_{75} và mAP lần lượt là 91.0%, 80.8%, 73.6%, tỏ ra vượt trội hoàn toàn so với các phương pháp còn lại. Xếp sau là hai phương pháp cũng đạt hiệu suất không hề thấp, rất sít sao với phương pháp trên. Double-Head R-CNN đạt kết quả thấp nhất với AP_{50} đạt 88.7%, AP_{75} đạt 78.4% và mAP đạt 71.0%. Tiếp theo, chúng tôi đi vào phân tích cụ thể các lớp trong các trường hợp tốt nhất và xấu nhất. Rõ ràng ta thấy, tất cả các lớp của bộ dữ liệu được thực nghiệm trên phương pháp Guided Anchoring đạt độ chính xác cao nhất so với phần còn lại với độ đo AP của lớp 'Table', 'Figure', 'Caption', 'Formula' đạt lần lượt là 92.7%, 81.6%, 73.3%, 46.9%. Trong khi đó, cũng với những class trên, Double-Head R-CNN chỉ đạt 91.5%, 80.6%,65.6% và 46.3% với độ đo AP, rõ ràng là cách biệt không nhỏ so với phương pháp tốt nhất hiện giờ. Kết quả của phương pháp Double-Head R-CNN không phải quá tệ, tuy nhiên nếu so sánh kiến trúc hai head riêng biệt với kiến trúc phân tầng và điều hướng các anchor linh hoạt thì rõ ràng Guided Anchoring đã mang lại những sự cải thiện đáng kể. Hơn nữa, các đặc trưng trong quá trình huấn luyện cũng được sàng lọc lại rất kỹ càng và chất lượng của các anchor cũng được đảm bảo, do đó kết quả của Guided Anchoring thực sự rất thuyết phục. Các kết ban đầu trên mô hình mặc định cho thấy Guided Anchoring với baseline là Faster R-CNN tương tự như Double Head hay Libra đã cho thấy hiệu suất vượt trội, do đó chúng tôi đã đề xuất Cascade R-CNN kết hợp với Guided Anchoring như đã đề cập tại phần 3.3. Kết quả cho thấy mô hình đề xuất của chúng tôi đã mang lại hiệu quả vượt trội với mAP lên đến 76.6%, kết quả này cao hơn 2.1% so với baseline được công bố bởi Dieu et al. khi sử dụng CascadeTabNet kết hợp với Fused Loss chỉ đạt được 74.5%.

Chương 5. **Kết luận và hướng phát triển**

Nhận thấy rằng nhu cầu trích xuất thông tin, phân tích dữ liệu document ngày càng tăng lên, chúng tôi đã tiến hành những nghiên cứu chuyên sâu để thực hiện bài toán object detection trên bộ dữ liệu tài liệu tiếng Việt dạng ảnh UIT-DODV. Sau khi

thực nghiệm trên các mô hình state-of-the-art như Double-Head R-CNN , Libra R-CNN, Guided Anchoring, chúng tôi ghi nhận được kết quả cao nhất với Guided Anchoring là 73.6% trên độ đo mAP. Với tiền đề trên, chúng tôi đề xuất mô hình phát hiện đối tượng Guided Anchoring Cascade R-CNN với sự kết hợp của hai phương pháp gồm Guided Anchoring và Cascade R-CNN. Kết quả của mô hình đề xuất của chúng tôi đã đạt được lên đến 76.6% với độ đo mAP, cao hơn mô hình baseline trên bộ dữ liệu UIT-DODV tới 2.1%.

Hướng tới những nghiên cứu sắp tới, chúng tôi sẽ không ngừng tìm hiểu và ứng dụng linh hoạt hơn các phương pháp phổ biến và mới mẻ trong bài toán phát hiện đối tượng. Hơn nữa, việc cải thiện mô hình về tốc độ huấn luyện và độ chính xác sẽ luôn được quan tâm và tập trung giải quyết, đồng thời sẽ tiến hành các nghiên cứu mở rộng và chuyên sâu để tạo tiền đề và nền tảng tốt cho những phương pháp sau này. Chúng tôi hy vọng là sự kết hợp mà chúng tôi đã đề xuất sẽ tạo ra những bước tiến lớn hơn cho những nghiên cứu trong tương lai hoặc ít nhất là đóng góp để làm tư liệu phát triển cho nền nghiên cứu trong lĩnh vực object detection trong tương lai.

SOURCE CODE THAM KHẢO:

Guided Anchoring Cascade RCNN

LỜI CẢM ƠN

Đồ án này được thực hiện với sự hướng dẫn của thầy Mai Tiến Dũng. Sau một học kì vô cùng thú vị và bổ ích, chúng em muốn dành một lời cảm ơn sâu sắc đến thầy, những chia sẻ và kinh nghiệm vô giá mà thầy đã mang đến trong học kì vừa qua sẽ giúp chúng em chuẩn bị tốt hơn những hành trang để tiếp tục trên con đường học tập cũng như nghiên cứu. Nhóm chúng em chúc thầy luôn có nhiều sức khỏe và niềm vui trong cuộc sống, đầy năng lượng để tiếp tục truyền những ngọn lửa đam mê đến những sinh viên của trường ta.

TÀI LIỆU THAM KHẢO

[1] Jwalin Bhatt, Khurram Azeem Hashmi, Muhammad Zeshan Afzal, and Didier Stricker. “A Survey of Graphical Page Object Detection with Deep Neural Networks”. In: Applied Sciences 11.12 (2021). ISSN: 2076-3417. DOI: 10.3390/app11125344. URL: https://www.mdpi.com/2076-3417/11/12/5344.

[2] Zhaowei Cai and Nuno Vasconcelos. “Cascade r-cnn: Delving into high quality object detection”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 6154–6162.

[3] Kai Chen et al. “MMDetection: Open MMLab Detection Toolbox and Benchmark”. In: arXiv preprint arXiv:1906.07155 (2019).

[4] Qiang Chen et al. “You only look one-level feature”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 13039– 13048.

[5] Linh Truong Dieu, Thuan Trong Nguyen, Nguyen D. Vo, Tam V. Nguyen, and Khang Nguyen. “Parsing Digitized Vietnamese Paper Documents”. In: Computer Analysis of Images and Patterns. Cham: Springer International Publishing, 2021, pp. 382–392.

[6] Kaiwen Duan et al. “Centernet: Keypoint triplets for object detection”. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 6569–6578.

[7] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 2117–2125

[8] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: European conference on computer vision. Springer. 2014, pp. 740–755.

[9] Duong Phi Long, Nguyen Trung Hieu, Nguyen Thanh Tuong Vi, Vo Duy Nguyen, and Nguyen Tan Tran Minh Khang. “Phat hien bang trong tai

liệu dạng anh su dung phuong phap đinh vi góc CornerNet". In: Proceedings of Fundamental and Applied Information Technology Research (FAIR). 2020.

[10] Thuan Trong Nguyen et al. "CDeRSNet: Towards High Performance Object Detection in Vietnamese Documents Images". In: International Conference on Multimedia Modelling (MMM). 2022.

[11] Jiangmiao Pang et al. "Libra r-cnn: Towards balanced learning for object detection". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 821–830.

[12] Bui Hai Phong, Thang Manh Hoang, and Thi-Lan Le. "An end-to-end framework for the detection of mathematical expressions in scientific document images". In: Expert Systems (2021), e12800.

[13] Heqian Qiu et al. "CrossDet: Crossline Representation for Object Detection". In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, pp. 3195–3204.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. 2016. arXiv: 1506. 01497 [cs.CV].

[15] Ningning Sun, Yuanping Zhu, and Xiaoming Hu. "Table Detection Using Boundary Refining via Corner Locating". In: Pattern Recognition and Computer Vision. Ed. by Zhouchen Lin et al. Cham: Springer International Publishing, 2019, pp. 135–146. ISBN: 978-3-030- 31654-9.

[16] Yunong Tian et al. "Apple detection during different growth stages in orchards using the improved YOLOV3 model". In: Computers and electronics in agriculture 157 (2019), pp. 417–426.

[17] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. "Fcos: Fully convolutional one-stage object detection". In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, pp. 9627–9636.

[18] Nguyen D Vo, Khanh Nguyen, Tam V Nguyen, and Khang Nguyen. “Ensemble of deep object detectors for page object detection”. In: Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication. 2018, pp. 1–6.

[19] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. “Region proposal by guided anchoring”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, pp. 2965–2974.

[20] Dihua Wu, Shuaichao Lv, Mei Jiang, and Huaibo Song. “Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments”. In: Computers and Electronics in Agriculture 178 (2020), p. 105742.

[21] Yue Wu et al. “Rethinking classification and localization for object detection”. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, pp. 10186–10195.

[22] Junaid Younas et al. “FFD: Figure and formula detection from document images”. In: 2019 Digital Image Computing: Techniques and Applications (DICTA). IEEE. 2019, pp. 1–7.

----------Hết----------