**CÔNG VIỆC 6: XÂY DỰNG MÔ ĐUN PHÂN TÍCH CẤU TRÚC VĂN BẢN**

# **LỜI NÓI ĐẦU**

Phân tích bộ cục tài liệu là một bước quan trọng trong quá trình xử lý tài liệu số, đặc biệt là trong các ứng dụng nhận dạng ký tự quang học(OCR), số hoá tài liệu và tự động hóa văn phòng. Công việc này liên quan đến việc chia tài liệu dưới dạng hình ảnh thành các vùng khác nhau, mỗi vùng tương ứng với một phần tử cụ thể trong tài liệu như văn bản, tiêu đề, bảng biểu, hình ảnh, danh sách, và các phần tử khác. Việc xác định chính xác vị trí và loại của từng phần tử trong tài liệu là cần thiết để trích xuất thông tin một cách hiệu quả và duy trì cấu trúc ban đầu của tài liệu.

Việc cân bằng giữa độ chính xác và hiệu suất là một vấn đề mang nhiều thách thức trong phân tích bố cục tài liệu. Thuật toán phân tích bố cục tài liệu dựa trên mô hình nhẹ PP-Picodet của PaddleDetection, một nền tảng mạnh mẽ trong lĩnh vực nhận dạng và phân tích hình ảnh. PP-PicoDet không chỉ giúp tối ưu hóa quá trình phân tích bố cục tài liệu mà còn mang đến những cải tiến vượt trội so với các phương pháp truyền thống. Với khả năng phát hiện các vùng bố cục chính xác cao và tốc độ xử lý nhanh, PP-PicoDet giúp giảm thiểu đáng kể thời gian và tài nguyên cần thiết để xử lý một lượng lớn tài liệu. Điều này đặc biệt quan trọng trong các ứng dụng yêu cầu xử lý thời gian thực hoặc khi phải xử lý trên các thiết bị di động có tài nguyên hạn chế

Với những ưu điểm nổi bật như vậy, PP-PicoDet đã trở thành một công cụ hữu ích và đáng tin cậy trong việc phân tích bố cục tài liệu, góp phần quan trọng vào quá trình số hóa tài liệu và tự động hóa văn phòng. Những cải tiến này không chỉ giúp nâng cao hiệu quả xử lý tài liệu mà còn mở ra nhiều cơ hội mới trong việc áp dụng công nghệ vào các lĩnh vực khác nhau, từ quản lý thông tin đến nghiên cứu và phát triển sản phẩm

# **MỤC LỤC**

[*LỜI NÓI ĐẦU 2*](#_q1qp4uqhbrtc)

[*MỤC LỤC 3*](#_f72r3y7v3844)

[*DANH MỤC HÌNH ẢNH 4*](#_8zifhbp8bmnt)

[*DANH MỤC BẢNG BIỂU 5*](#_e83wvp34ybbu)

[*DANH MỤC CHỮ VIẾT TẮT 6*](#_q9bqhfaq9nvl)

[*CHƯƠNG 1: TỔNG QUAN VỀ PHÂN TÍCH CẤU TRÚC VĂN BẢN 7*](#_hvhfav5bfk8z)

[*CHƯƠNG 2: MÔ HÌNH PP-Picodet 12*](#_yv1fx0gv4waj)

[*2.1 Backbone(LCNet) 13*](#_jo3vkv63ij6)

[*2.2 Neural Architecture Search. 14*](#_2q3kamcs3pwk)

[*2.3 CSP-PAN and Detector Head 15*](#_wij0cs711yv)

[*CHƯƠNG 3: THUẬT TOÁN FGD 17*](#_tgjrid7kez0l)

[*CHƯƠNG 4: THỬ NGHIỆM VÀ ĐÁNH GIÁ 19*](#_tjsppef0ak9u)

[*KẾT LUẬN 21*](#_mouqax7lnlcd)

[*TÀI LIỆU THAM KHẢO 22*](#_vvjtrsiapilh)

# 

# **DANH MỤC HÌNH ẢNH**

[Hình 1.Kiến trúc mô hình YOLO 7](#_Toc176560614)

[*Hình 2.Dự đoán vị trí Anchor box trong Yolo 8*](#_Toc176560615)

[*Hình 3.Cấu trúc mô hình FCOS 8*](#_Toc176560616)

[*Hình 4.Ví dụ về phân tích cấu trúc tài liệu 9*](#_Toc176560617)

[*Hình 5.Thử nghiệm và so sánh mAP của các mô hình nhẹ khác nhau trên MS-COCO 10*](#_Toc176560618)

[*Hình 6.Kiến trúc của PP-Picodet 12*](#_Toc176560619)

[*Hình 7.Kiến trúc ES 13*](#_Toc176560620)

[*Hình 8.Cấu trúc CSP-PAN 16*](#_Toc176560621)

[*Hình 9.Minh họa về FGD vao gồm focal distillation và global distillation 17*](#_Toc176560622)

# **DANH MỤC BẢNG BIỂU**

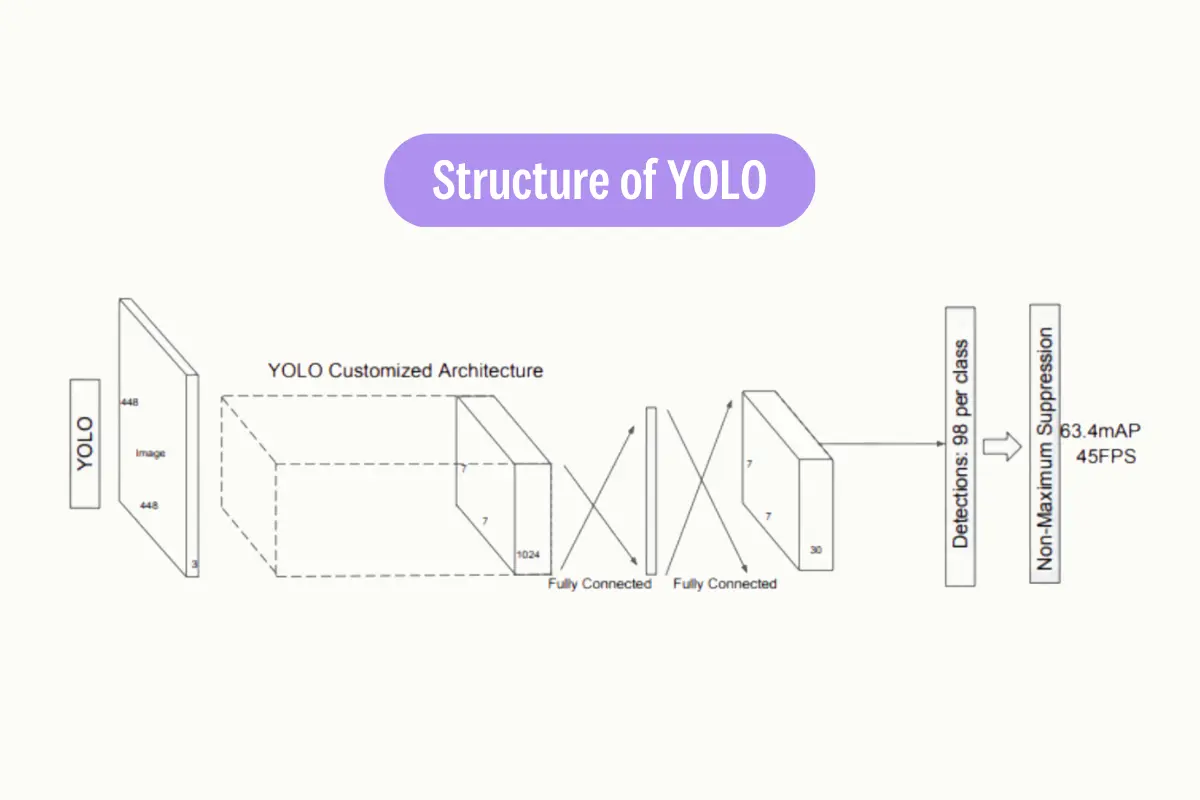
[*Bảng 1.Các thí nghiệm phân tích cấu trúc tài liệu trên PubLayNet 20*](#_Toc176560525)

# **DANH MỤC CHỮ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
| **STT** | **Từ viết tắt** | **Từ viết tắt đầy đủ** |
| 1 | CSP | Cross-Stage Partial Network |
| 2 | EA | Evolutionary Algorithm |
| 3 | FGD | Focal and Global Knowledge Distillation |
| 4 | mAP | mean Average Precision |
| 5 | NAS | Neural Architecture Search |
| 6 | OCR | Optical Character Recognition |
| 7 | SE | Squeeze-and-Excitation |
| 8 | ESNet | Enhanced ShuffleNet |

# **CHƯƠNG 1: TỔNG QUAN VỀ PHÂN TÍCH CẤU TRÚC VĂN BẢN**

Nhận diện đối tượng( Object detection) được ứng dụng rộng rãi trong nhiều nhiệm vụ của thị giác máy tính, bao gồm lái xe tự động, giao thông thông minh, kiểm tra chất lượng công nghiệp, theo dõi đối tượng, và nhiều lĩnh vực khác. Các mô hình hai giai đoạn(two-stage models) thương mang lại hiệu suất cao hơn. Tuy nhiên, loại kiến trúc mạng này tiêu tốn nhiều tài nguyên, khiến việc áp dụng vào các ứng dụng thực tế trở nên khó khăn. Để giải quyết vấn đề này, các bộ phát hiện đối tượng nhẹ dành cho các thiết bị di động đã thu hút sự quan tâm của các nhà nghiên cứu, nhằm mục tiêu thiết kế các mô hình nhận diện có hiệu suất cao.

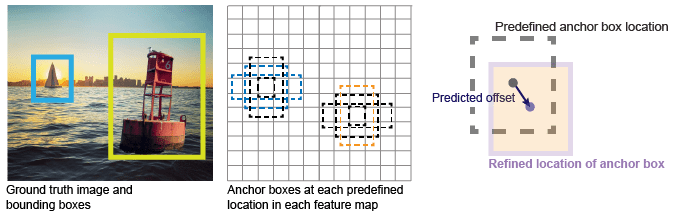


Hình 1.Kiến trúc mô hình YOLO

Các thuật toán phát hiện đối tượng hiện đại thuộc họ YOLO[1,2,3,4] đã trở nên phổ biến vì chúng quan tâm đến vấn đề hạn chế về tài nguyên trong quá trình thực thi. So với các mô hình hai giai đoạn, họ YOLO có hiệu suất tốt hơn và độ chính xác cao hơn. Tuy nhiên, họ YOLO vẫn gặp phải một số vấn đề như:

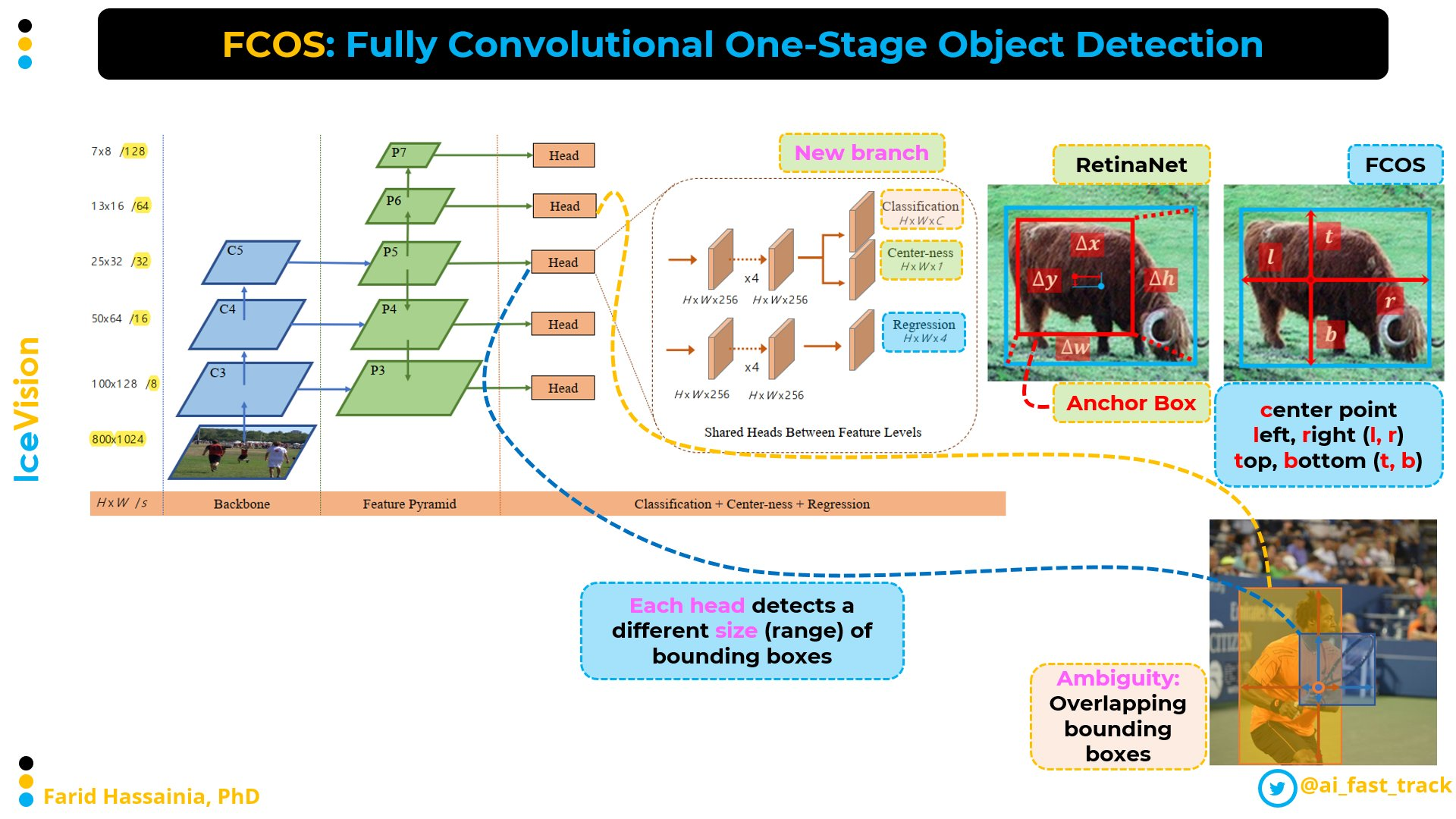
(1)Cần phải thiết kế lại hộp neo( anchor boxes) một cách thủ công để phù hợp với các tập dữ liệu khác nhau.

(2)Vấn đề mất cân bằng giữa mẫu dương và mẫu âm, do đa phần các hộp neo được tạo ra đều là mẫu âm.



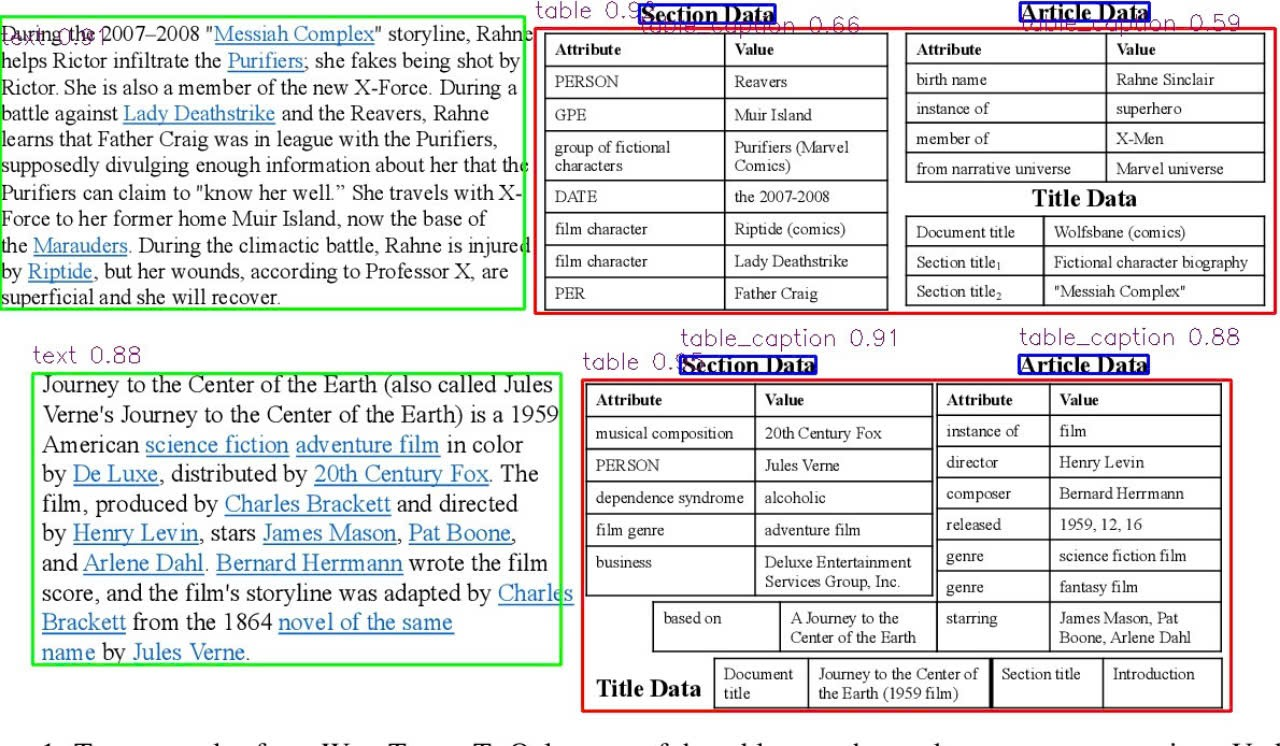
Hình 2.Dự đoán vị trí Anchor box trong Yolo

Trong những năm gần đây, với sự phát triển vượt bậc của thị giác máy tính, nhiều công trình nghiên cứu đã được thực hiện nhằm phát triển các kiến trúc bộ nhận diện đối tượng hiệu quả hơn. Một trong những hướng đi nổi bật là sự ra đời của các bộ nhận diện không cần anchor (anchor-free detectors). Đây là những mô hình không cần sử dụng các hộp neo (anchor boxes) để xác định các đối tượng trong hình ảnh, giúp đơn giản hóa quá trình nhận diện và giảm thiểu các vấn đề liên quan đến thiết kế thủ công và tối ưu hóa các tham số. Một ví dụ điển hình trong số này là FCOS[5], một bộ nhận diện không cần anchor đã giải quyết được vấn đề chồng lấn trong các nhãn ground-truth, tức là các nhãn thực tế của đối tượng trong dữ liệu huấn luyện.



Hình 3.Cấu trúc mô hình FCOS

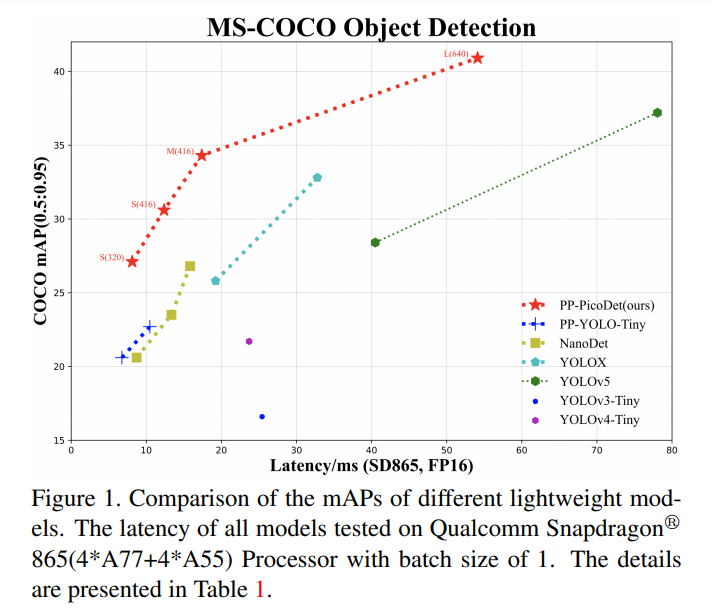
Điểm nổi bật của FCOS so với các mô hình không cần anchor khác là nó không yêu cầu tinh chỉnh các siêu tham số phức tạp, điều này giúp giảm bớt khối lượng công việc khi triển khai mô hình vào các ứng dụng thực tế. Tuy nhiên, cần lưu ý rằng hầu hết các bộ nhận diện không cần anchor hiện nay đều được thiết kế để hoạt động trên các hệ thống máy chủ lớn, đòi hỏi nhiều tài nguyên tính toán. Trong số rất ít các bộ nhận diện không cần anchor dành cho thiết bị di động, có thể kể đến NanoDet[6] và YOLOX-Nano[4]. Đây là những bộ nhận diện nhỏ gọn, hướng đến các ứng dụng trên các thiết bị có tài nguyên hạn chế. Tuy nhiên, một thách thức lớn mà các bộ nhận diện nhẹ không cần anchor đang đối mặt là khả năng cân bằng giữa độ chính xác và hiệu suất. Các mô hình nhẹ thường khó đạt được cả hai yếu tố này, dẫn đến việc phải đánh đổi giữa một trong hai.



Hình 4.Ví dụ về phân tích cấu trúc tài liệu

Phân tích cấu trúc văn bản không chỉ liên quan đến việc nhận diện các đối tượng trong hình ảnh tài liệu mà còn tập trung vào việc xác định và phân loại các vùng quan trọng, chẳng hạn như đoạn văn, tiêu đề, hình ảnh, bảng biểu,công thức toán học, và các phần khác bên trong một tài liệu. Việc nhận diện đúng cấu trúc của tài liệu giúp cho các hệ thống OCR không chỉ trích xuất được nội dung chữ viết mà còn hiểu được nghĩa và trình tự logic của tài liệu. Thách thức trong phân tích cấu trúc văn bản là làm thế nào để cân bằng giữa việc xử lý hiệu quả các vùng tài liệu phức tạp mà vẫn duy trì được độ chính xác cao trong việc phân loại và nhận diện. Các bộ nhận diện nhẹ không cần anchor có thể đóng vai trò quan trọng trong việc giải quyết vấn đề này, nhưng cũng đòi hỏi sự cải tiến và tối ưu hóa hơn nữa để đáp ứng được yêu cầu của các ứng dụng thực tế.

PP-Structure - một sản phẩm của PaddlePaddle(PArallel Distributed Deep Learning) là một trong những nền tảng mã nguồn mở hàng đầu về nhận diện ký tự quang học (OCR), không chỉ hỗ trợ nhận diện văn bản mà còn tích hợp các module phân tích bố cục mạnh mẽ. PP-Structure đã phát triển nhiều giải pháp tiên tiến để xử lý và phân tích cấu trúc văn bản trong tài liệu, giúp nhận diện và phân loại chính xác các vùng quan trọng như văn bản, tiêu đề, bảng biểu, hình ảnh, công thức toán học và nhiều thành phần khác.Trong PP-Structure, PaddlePaddle đã sử dụng thuật toán phát hiện đối tượng YOLOv2[7] làm bộ phát hiện bố cục tài liệu. Trong PP-StructureV2, họ đã sử dụng bộ phát hiện nhẹ hơn là PP-Picodet[8], PP-Picodet đạt hiệu suất vượt trội trên các thiết bị di động. Ngoài ra, trong phiên bản này, Thuật toán FGD[9] còn được sử dụng để cải thiện độ chính xác của mô hình.



Hình 5.Thử nghiệm và so sánh mAP của các mô hình nhẹ khác nhau trên MS-COCO

Hình trên là biểu đồ so sánh độ chính xác của các mô hình phát hiện đối tượng trên tập dữ liệu COCO. PicoDet-S với chỉ 0.99 triệu tham số đạt được 30.6% mAP, cải thiện 4.8% mAP so với YOLOX-Nano và giảm độ trễ khi thực thi trên CPU di động xuống 55%. So với NanoDet, PicoDet-S cải thiện mAP thêm 7.1%. Mô hình này đạt tốc độ 123 FPS (150 FPS khi sử dụng Paddle Lite) trên CPU ARM di động với kích thước đầu vào là 320. PicoDet-L với chỉ 3.3 triệu tham số đạt được 40.9% mAP, cải thiện 3.7% mAP và nhanh hơn 44% so với YOLOv5s.

Nhìn chung, các mô hình PicoDet không chỉ vượt trội so với các mô hình hiện tại về độ chính xác mà còn tối ưu hóa hiệu suất trên thiết bị di động. Sự cải thiện đáng kể về mAP và giảm thiểu độ trễ xử lý khẳng định sự thành công của PicoDet trong việc đáp ứng các yêu cầu của phát hiện đối tượng nhẹ. Các kết quả này cung cấp một giải pháp mạnh mẽ cho các ứng dụng thực tế cần hiệu suất cao và tính toán hiệu quả trên các thiết bị di động.

# 

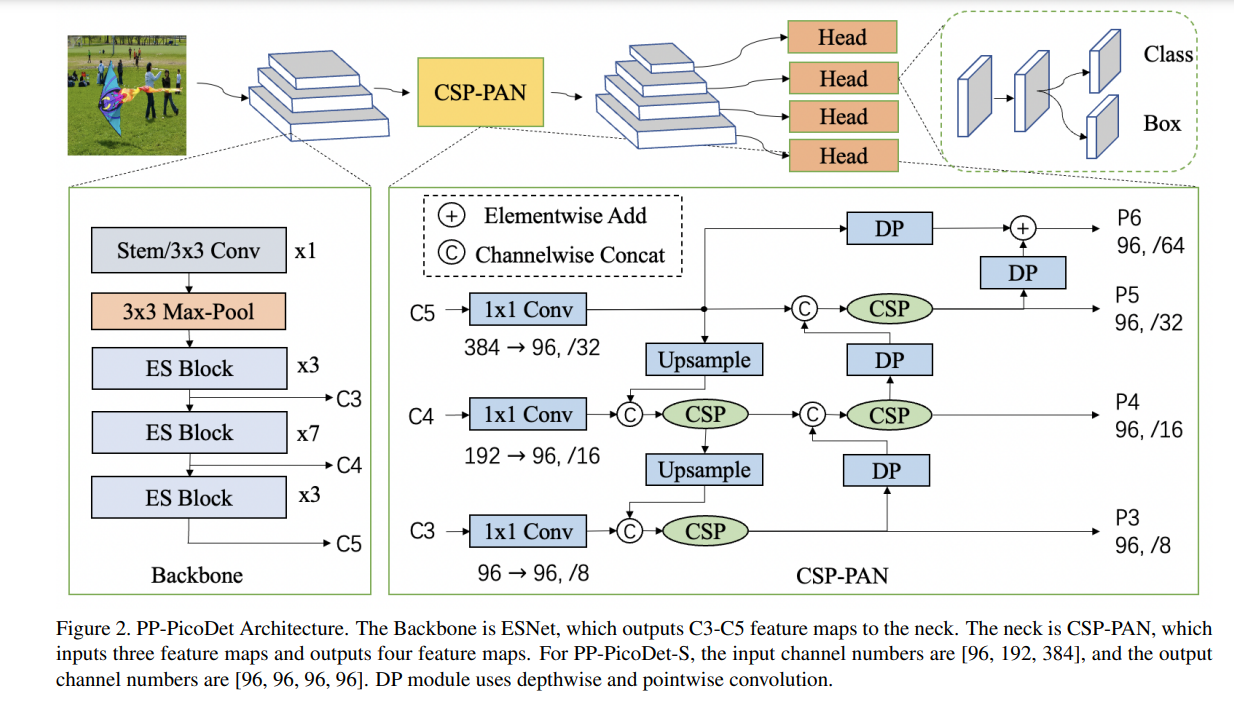
# **CHƯƠNG 2: MÔ HÌNH PP-Picodet**

Để có thể dễ dàng triển khai trên các thiết bị di động và đạt được độ chính xác cao. PaddlePaddle đã phát triển một bộ phát hiện có tên là PP-Picodet. Mục tiêu là tạo ra một mô hình không chỉ đạt được hiệu suất nhanh chóng trên các thiết bị bị có tài nguyên hạn chế, mà còn duy trì được độ chính xác cao trong việc phát hiện các đối tượng. Những đóng góp chính của PP-Picodet bao gồm:

-Áp dụng cấu trúc CSP để xây dựng CSP-PAN làm phần cổ(neck) của mạng. CSP-PAN thống nhất số kênh đầu vào bằng phép tích chập 1x1 cho tất cả các nhánh của phần cổ, điều này giúp tăng cường khả năng trích xuất đặc trưng và làm giảm số lượng tham số của mạng. Đặc biệt, phép tích chập 5x5 cũng được thay thế cho phép tích chập theo độ sâu 3x3 để mở rộng trường quan sát cho mạng.

-Đề xuất và sử dụng một backbone mới, gọi là Enhanced ShuffleNet(ESNET)có hiệu suất tốt hơn so với mạng ShuffleNetV2.

-Đề xuất quy trình cải tiến và sử dụng One-Shot Neural Architecture(NAS) để tự động tìm kiếm kiến trúc tối ưu cho việc phát hiện đối tượng. Các mô hình được tạo ra tự NAS đạt được sự cân bằng tốt hơn giữa hiệu suất và độ chính xác.



Hình 6.Kiến trúc của PP-Picodet

Hình trên minh hoạ kiến trúc của PP-Picodet với Backbone là ESNet, đầu ra là các feature maps C3-C5 đến phần neck. Phần neck là CSP-PAN, nó nhận 3 feature map đầu vào và tạo ra bốn feature map đầu ra. Đối với PP-Picodet-S, số kênh đầu vào là [96,192,384], và số lượng kênh đầu ra là [96,96,96,96]. Module DP sử dụng các phép tích chập theo chiều sâu( Depthwise convolution) và điểm( Pointwise convolution)

Thông qua các tối ưu trong PP-Picodet đã thu được kết quả vượt trội hơn hẳn so với các kết quả hiện tại trong lĩnh vực phát hiện các đối tượng nhẹ. Mô hình đã đạt được hiệu suất cao với số lượng tham số và tính toán thấp, đáp ứng yêu cầu của các hệ thống triển khai khác nhau.

## **2.1 Backbone(LCNet)**

Dựa trên nhiều thí nghiệm, ShuffleNetV2 được chứng minh là hoạt động ổn định hơn so với các mạng khác trên các thiết bị di động. Để cải thiện hiệu suất của ShuffleNetV2, nhóm tác giả của PaddlePaddle đã áp dụng một số phương pháp từ PPLCNet để nâng cấp trúc mạng này và xây dựng một backbone mới, được gọi là Enhanced ShuffleNet(ESNet). Hình dưới đây mô tả chi tiết về ES Block của ESNet.

|  |  |
| --- | --- |
| 1. *stride=2* | 1. *stride=1* |

Hình 7.Kiến trúc ES

Module SE đã chứng minh khả năng vượt trội trong việc gán trọng số cho các kênh mạng để làm nổi bật các đặc trưng quan trọng. Vì vậy, module SE đã được thêm vào tất cả các khối mạng. Tương tự như MobileNetV3, các hàm kích hoạt được sử dụng trong hai lớp của module SE lần lượt là ReLU và H-Sigmoid.

Channel shuffle trong ShuffleNetV2 giúp trao đổi thông tin giữa các kênh mạng, nhưng đồng thời cũng gây ra sự mất mát của các đặc trưng. Để giải quyết vấn đề này, việc thêm các lớp tích chập tách biệt theo chiều sâu( depthwise convolution) và tích chập theo điểm( pointwise convolution) để kết hợp thông tin từ các kênh khác nhau khi stride là 2(Hình 7a). Tác giả của GhostNet đã đề xuất một module mới được gọi là Shost module, có thể tạo ra nhiều feature map hơn với số lượng tham số ít hơn, từ đó cải thiện khả năng học của mạng. Vì vậy Ghost module được tích hợp vào các khối stride bằng 1 để tăng cường hiệu suất của ESNet(Hình 7b).

## **2.2 Neural Architecture Search.**

Các module phát hiện đối tượng, khi được trang bị các backbone có hiệu suất cao trong phân loại, có thể không đạt được hiệu suất tối ưu do sự khác biệt giữa các tác vụ phân loại đối tượng. Vì vậy, thay vì tìm kiếm một bộ phân loại tốt hơn, PP-Picodet tiến hành huấn luyện và tìm kiếm trực tiếp trên mạng supernet, từ đó tiết kiệm đáng kể chi phí tính toán và tối ưu hóa cho nhiệm vụ phát hiện thay vì phân loại.

Khung hoạt động bao gồm 2 bước:

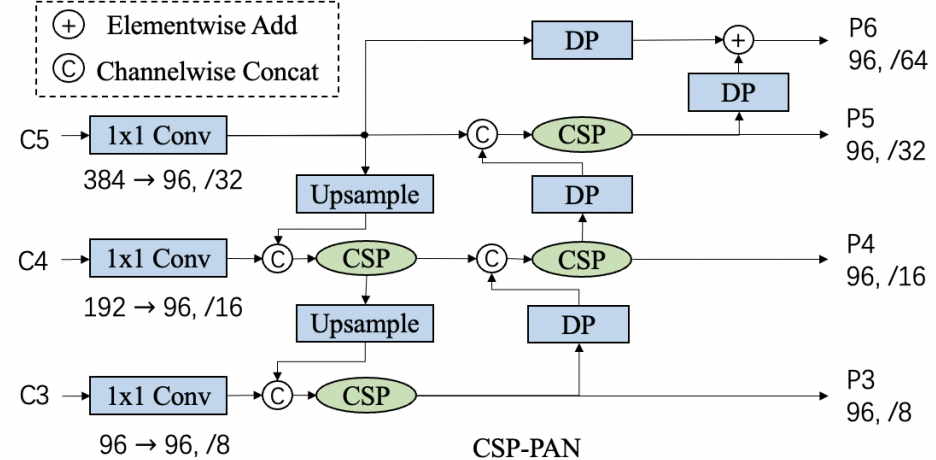
1. Huấn luyện supernet trên các tập dữ liệu phát hiện đối tượng
2. Tìm kiếm kiến trúc trên supernet đã huấn luyện bằng thuật toán EA

Để đơn giản hóa quá trình, phương pháp tìm kiếm theo kênh được sử dụng cho backbone. Cụ thể, các tùy chọn tỷ lệ linh hoạt được cung cấp để lựa chọn các tỷ lệ kênh khác nhau. Các tỷ lệ được chọn một cách ngẫu nhiên trong tập [0.5,0.675,0.75,0.857,1]. Ví dụ, tỷ lệ 0.5 tương ứng với việc thu nhỏ chiều rộng của mô hình xuống còn 0.5 so với mô hình đầy đủ. Do đó, thay vì có thể cải thiện tốc độ thực thi trên các thiết bị phần cứng.

## **2.3 CSP-PAN and Detector Head**

Cấu trúc PAN[10] được sử dụng để thu được các feature map ở nhiều mức độ khác nhau, trong khu cấu trúc CSP được áp dụng kết hợp và hợp nhất các đặc trưng giữa các feature map liền kề. Cấu trúc CSP được sử dụng rộng rãi trong phần neck của các mô hình YOLOv4 và YOLOX . Trong phiên bản CSP-PAN gốc, số lượng kênh trong mỗi feature map đầu ra được giữ nguyên so với đầu vào từ backbone. Tuy nhiên, các cấu trúc với số lượng kênh lớn thường có chi phí tính toán cao đối với các thiết bị di động. Đề giải quyết vấn đề này, số lượng kênh trong tất cả các feature map được làm bằng nhau với số lượng kênh nhỏ nhất thông qua phép toán convolution 1x1. Sau đó, việc kết hợp các đặc trưng theo chiều từ trên xuống và từ dưới lên được thực hiện thông qua cấu trúc CSP. Việc giảm kích thước các đặc trưng mang lại chi phí tính toán thấp hơn mà không làm giảm độ chính xác.

Ngoài ra, một feature map bổ sung được thêm vào trên đỉnh của CSP-PAN để phát hiện nhiều đối tượng hơn. Đồng thời, tất cả các phép toán convolution ngoài phép toán convolution 1 × 1 đều là convolution tách biệt theo chiều sâu. Convolution tách biệt theo chiều sâu mở rộng trường tiếp nhận thông qua phép toán convolution 5 × 5. Cấu trúc này mang lại sự gia tăng đáng kể về độ chính xác với số lượng tham số ít hơn nhiều. Cấu trúc cụ thể được trình bày trong hình dưới đây:



Hình 8.Cấu trúc CSP-PAN

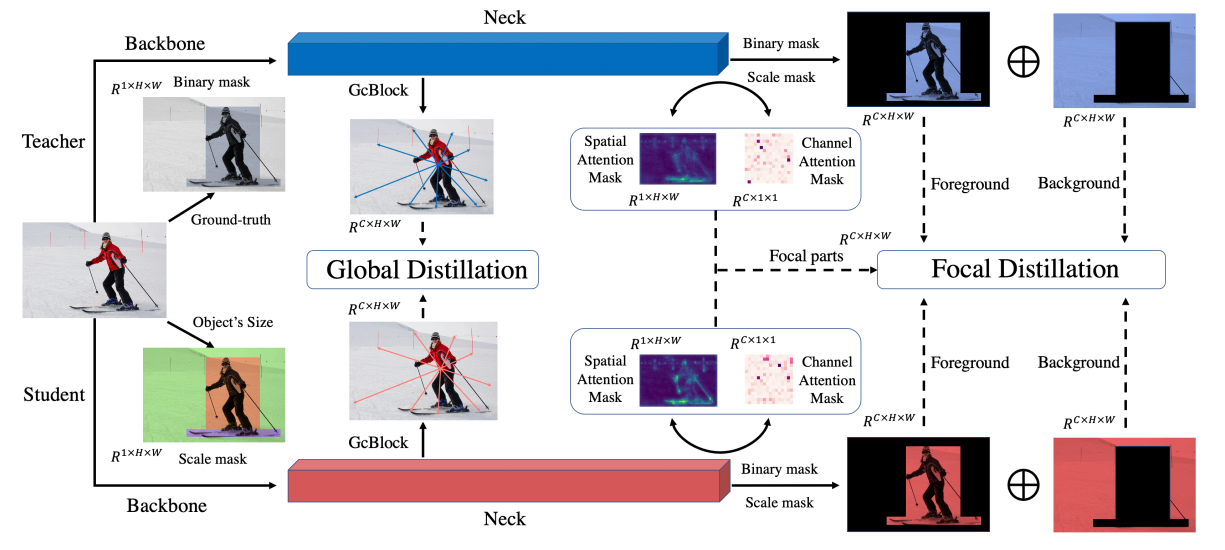
Trong Detector head, phép toán convolution tách biệt theo chiều sâu và phép toán convolution 5x5 được sử dụng để mở rộng receptive field. Số lượng phép toán convolution tách biệt theo chiều sâu có thể đặt thành 2,4 hoặc nhiều hơn. Cấu trúc tổng thể của mạng được thể hiện trong Hình 6. Cả phần neck và Detector head đều có bốn nhánh. Số lượng kênh trong Detector head được giữ nguyên so với module neck và các nhánh phân loại và hồi quy.

# **CHƯƠNG 3: THUẬT TOÁN FGD**

FGD là một thuật toán tinh chế kiến thức được áp dụng trong lĩnh vực phát hiện đối tượng, được giới thiệu bởi Yang et al, vào năm 2022. Thuật toán này kết hợp hai phương pháp: Tinh chế tập trung( focal distillation) và tinh chế toàn cục(global distillation)

Tinh chế tập trung: Phương pháp này tập trung vào việc phân tách các vùng foreground (tiền cảnh) và background (hậu cảnh) của hình ảnh. Quá trình này buộc mô hình học sinh (student model) phải chú ý đặc biệt đến các pixel và kênh quan trọng trong bản đồ đặc trưng của mô hình thầy (teacher model). Tinh chế tập trung giúp mô hình học sinh học được những đặc điểm nổi bật và quan trọng của hình ảnh từ mô hình thầy.

Tinh chế toàn cục: Phương pháp này nhằm tái thiết lập mối quan hệ toàn cục giữa các pixel khác nhau và chuyển giao thông tin từ mô hình thầy sang mô hình học sinh. Tinh chế toàn cục bù đắp cho sự thiếu hụt thông tin toàn cục mà tinh chế tập trung có thể gặp phải, từ đó cải thiện khả năng nhận diện và hiểu biết của mô hình học sinh về ngữ cảnh toàn cục của dữ liệu.

****

Hình 9.Minh họa về FGD vao gồm focal distillation và global distillation

Có thể thấy được rằng FGD không chỉ phân tách tiền cảnh và hậu cảnh mà còn giúp student network chú ý tốt hơn đến thông tin quan trọng trong feature map của mạng teacher. Global distillation nối khoảng cách giữa ngữ cảnh của student network và teacher network.

Trong lĩnh vực phân tích cấu trúc tài liệu, việc áp dụng FGD mang lại nhiều lợi ích đáng kể. Focal distillation tập trung vào việc giúp model student chú ý đến các phần tử quan trọng trong tài liệu, như tiêu đề, đoạn văn, và các chi tiết cấu trúc chính. Bằng cách điều chỉnh sự chú ý của model student vào các đặc điểm quan trọng từ mô hình thầy, FGD giúp nâng cao khả năng phân loại và phân tách các vùng chức năng trong tài liệu một cách chính xác hơn. Điều này đảm bảo rằng các phần tử quan trọng được nhận diện rõ ràng và phân tích hiệu quả, làm tăng chất lượng của việc phân tích cấu trúc tài liệu. Đồng thời, global distillation giúp duy trì ngữ cảnh toàn cục của tài liệu. Phương pháp này đảm bảo rằng model student không chỉ học được các chi tiết cụ thể mà còn hiểu được mối liên hệ tổng thể giữa các phần khác nhau của tài liệu. Bằng cách chuyển giao thông tin toàn cục từ model teacher, FGD giúp model student nắm bắt và giữ gìn mối quan hệ giữa các phần tử trong tài liệu. Điều này rất quan trọng trong phân tích cấu trúc tài liệu, vì nó hỗ trợ mô hình hiểu cách các phần tử khác nhau liên kết và tương tác với nhau, từ đó cải thiện khả năng phân tích và hiểu biết về cấu trúc tổng thể của tài liệu.

Kết quả của việc áp dụng FGD cho mô hình phát hiện đối tượng PP-PicoDet đã cho thấy những cải thiện đáng kể trong cả độ chính xác và hiệu suất. Cụ thể, mô hình học sinh PP-PicoDet dựa trên LCNet1.0x đã đạt được sự cải thiện 0.5% về chỉ số mAP nhờ vào việc chuyển giao kiến thức từ mô hình thầy PP-PicoDet dựa trên LCNet2.5x. Điều này cho thấy rằng phương pháp FGD không chỉ giúp tăng cường khả năng nhận diện chi tiết mà còn duy trì sự cân bằng giữa độ chính xác và hiệu suất.

# **CHƯƠNG 4: THỬ NGHIỆM VÀ ĐÁNH GIÁ**

Để có thể thử nghiệm và đánh giá hiệu quả của mô hình PP-Picodet trong tác vụ phân tích bố cục văn bản, bộ dữ liệu tiêu biểu là PubLayNet được sử dụng trong quá trình huấn luyện và đánh giá mô hình. PubLayNet (Zhong, Tang, và Yepes 2019), một bộ dữ liệu quy mô lớn với 335.703 hình ảnh trong tập huấn luyện, 11.245 hình ảnh trong tập kiểm tra và 11.405 hình ảnh trong tập kiểm thử. Bộ dữ liệu này bao gồm các yếu tố bố cục tài liệu như văn bản, tiêu đề, danh sách, bảng và hình ảnh, giúp mô hình của chúng tôi có khả năng nhận diện và phân tích bố cục tài liệu một cách toàn diện. Chỉ số mAP (Mean Average Precision) được sử dụng để đánh giá hiệu suất của các mô hình.

Trong quá trình huấn luyện mô hình, Momentum được sử dụng với hệ số là 0.9 và trọng số giảm là 4e-5. Chiến lược điều chỉnh tỷ lệ học theo hàm Cosine được áp dụng với tỷ lệ học là 0.4. Kích thước lô(Batch size) và số lượt epoch được đặc lần lượt là 24 và 8.

Các thí nghiệm phân tích bố cục tài liệu được thực hiện trên bộ dữ liệu PubLayNet và kết quả được trình bày trong Bảng 1. PP-YOLOv2 được sử dụng để có thể đánh giá hiệu suất của PP-Picodet. Mặc dù PP-Picodet-LCNet2.5x thể hiện hiệu suất tốt hơn PP-YOLOv2, nhưng mAP của nó giảm 1,1%. Tuy nhiên, bằng cách điều chỉnh kích thước đầu vào, mAP có thể được cải thiện thêm 1,7%, cao hơn so với mức cơ bản. Để tạo ra một mô hình nhẹ, mô hình 1.0x được huấn luyện với FGD, sử dụng mô hình 2.5x trước đó làm mô hình cha. Kết quả cho thấy mAP cuối cùng vượt quá mức cơ bản 0,4%, với tốc độ thực thi tăng lên gấp 11 lần và kích thước lưu trữ mô hình giảm 95%.

|  |  |  |  |
| --- | --- | --- | --- |
| **Strategy** | **mAP(%)** | **Speed(ms)** | **Model Size(M)** |
| PP-YOLOv2(640\*640) | 93.6 | 51.2 | 221 |
| PP-PicoDet-LCNet2.5x(640\*640) | 92.5 | 53.2 | 29.7 |
| PP-PicoDet-LCNet2.5x(800\*608) | 94.2 | 83.1 | 29.7 |
| PP-PicoDet-LCNet1.0x(640\*640) | 93.5 | 41.2 | 9.7 |
| PP-PicoDet-LCNet1.0x(800\*608)+FGD | 94.0 | 41.2 | 9.7 |

Bảng 1.Các thí nghiệm phân tích cấu trúc tài liệu trên PubLayNet

Nhìn chung, các mô hình PP-PicoDet cho thấy sự cải thiện rõ rệt về kích thước mô hình và tốc độ thực thi so với PP-YOLOv2, trong khi vẫn duy trì hoặc cải thiện độ chính xác. Đặc biệt, mô hình PP-PicoDet-LCNet1.0x với FGD kết hợp lợi ích của giảm kích thước mô hình và tăng mAP, cho thấy đây là sự lựa chọn tối ưu cho các ứng dụng yêu cầu mô hình nhẹ và hiệu suất cao.

# **KẾT LUẬN**

Mô hình PP-PicoDet đã thể hiện những cải tiến vượt trội trong việc phân tích cấu trúc tài liệu so với các mô hình trước đây như PP-YOLOv2, nhờ vào khả năng tối ưu hóa kích thước mô hình và cải thiện hiệu suất. Đặc biệt, việc áp dụng FGD trong PP-PicoDet đã giúp giảm kích thước mô hình mà vẫn duy trì hoặc thậm chí nâng cao độ chính xác, điều này rất quan trọng trong các tác vụ OCR. Mô hình này cung cấp khả năng phân tích cấu trúc tài liệu một cách nhanh chóng và hiệu quả, hỗ trợ việc nhận diện và phân loại các thành phần của tài liệu với độ chính xác cao. Các thành phần nâng cao như SE và Ghost modules đã góp phần vào việc cải thiện khả năng học tập và xử lý của mô hình, giúp PP-PicoDet hoạt động tốt trong các bài toán phức tạp liên quan đến phân tích cấu trúc tài liệu.

Để nâng cao hiệu suất và mở rộng ứng dụng của PP-PicoDet trong phân tích cấu trúc tài liệu, các nghiên cứu và phát triển có thể tập trung vào một số hướng chính. Đầu tiên, việc cải tiến và tối ưu hóa mô hình để nâng cao độ chính xác trong các tình huống thực tế và trên các loại tài liệu khác nhau là điều cần thiết. Thứ hai, mở rộng ứng dụng của PP-PicoDet sang các lĩnh vực khác như phân tích tài liệu đa ngữ hoặc nhận diện văn bản trong các môi trường phức tạp có thể khai thác được nhiều lợi ích của mô hình hơn trong tác vụ xử lý ngôn ngữ. Và việc tối ưu hóa tài nguyên, cải thiện tốc độ thực thi của mô hình cũng sẽ là các yếu tố giúp tăng cường khả năng triển khai của PP-Picodet trong các ứng dụng OCR và phân tích tài liệu rộng hơn.

# **TÀI LIỆU THAM KHẢO**

*[1] Joseph Redmon et. Yolov3: An incremental improvement, 2018.*

*[2]Alexey Bochkovskiy et. Yolov4: Optimal speed and accuracy of object detection, 2020.*

*[3]Glenn Jocher et. ultralytics/yolov5: v6.0 - YOLOv5n ’Nano’ models, Roboflow integration, TensorFlow export, OpenCV DNN support, 2021.*

*[4] Zheng Ge et. Yolox: Exceeding yolo series, 2021.*

*[5]Zhi Tian et. Fcos: Fully convolutional one-stage object detection, 2021.*

*[6]NanoDet. 2021, https://github.com/ RangiLyu/nanodet.*

*[7] Xin Huang et. Pp-yolov2: A practical object detector, 2021.*

*[8]Guanghua Yu et. PP-PicoDet: A Better Real-Time Object Detector on Mobile Devices, 2021.*

*[9]Yang et. Focal and global knowledge distillation for detectors, 2022.*

*[10] Shu Liu et. Path aggregation network for instance segmentation, 2018.*