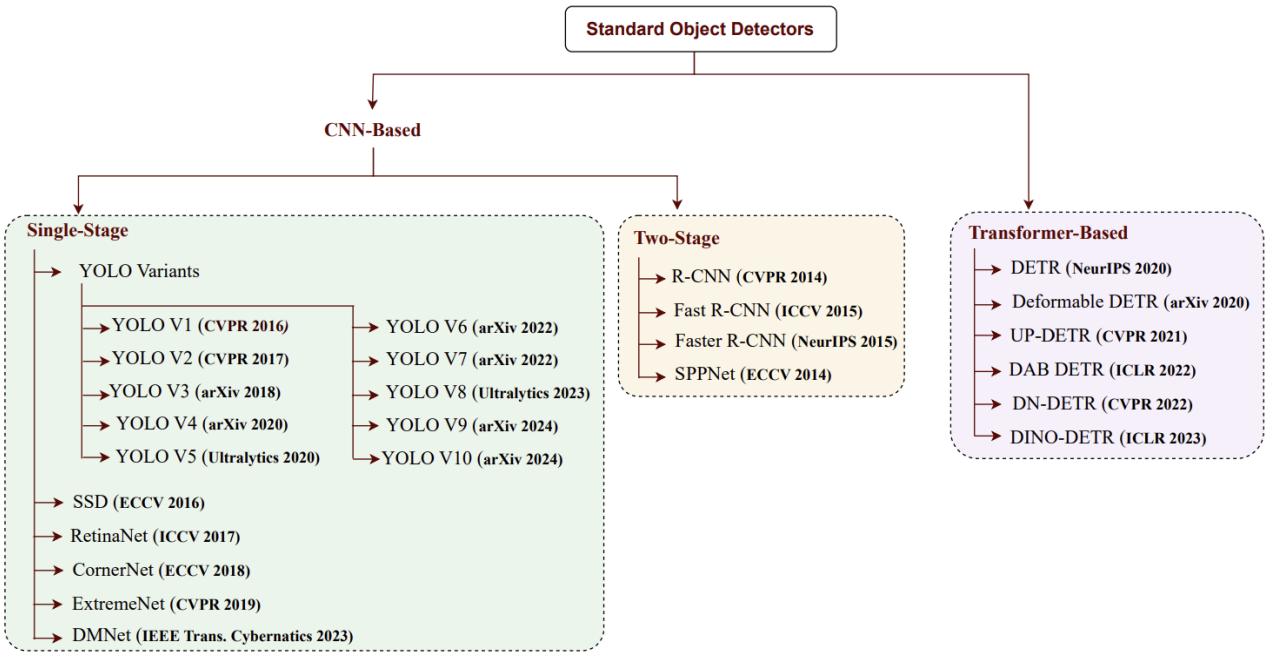
Report: Tổng quan về các mô hình nhận diện

1. Sơ qua một số mô hình phổ biến



Đặc điểm	YOLO v11	DMNet	DINO-DETR
Loại mô hình	CNN-based, Single-Stage	CNN-based, Single-Stage	Transformer-Based
Mục tiêu chính	Tốc độ và khả năng nhận diện thời gian thực	Độ chính xác trong môi trường phức tạp	Độ chính xác cao với transformer
Kiến trúc	Chuỗi kiến trúc YOLO, tối ưu hóa cho tốc độ	Mạng module biến dạng (deformable module network)	Dựa trên DETR, sử dụng transformer với cải tiến
Loại xử lý	Một giai đoạn (Single-Stage)	Một giai đoạn (Single-Stage)	Transformer-Based (Sử dụng transformer)
Cải tiến nổi bật	Tối ưu hóa các lớp mạng để tăng tốc độ suy luận	Các lớp convolution biến dạng đa cấp	Cơ chế khử nhiễu, truy vấn dày đặc, attention
Tốc độ xử lý	Rất cao, phù hợp với thời gian thực	Trung bình, chậm hơn YOLO V10	Chậm hơn (transformer yêu cầu tài nguyên lớn)
Độ chính xác nhận diện	Cao (cải thiện so với các phiên bản YOLO trước đó)	Rất cao (tối ưu cho các cấu trúc phức tạp)	Rất cao (độ chính xác và khả năng localization tốt)
Phương pháp trích xuất đặc trưng	Mạng CNN (Convolutional Neural Network)	Các lớp convolution biến dạng	Transformer với attention
Xử lý che khuất và vật thể nhỏ	Trung bình (hạn chế bởi lưới grid của YOLO)	Tốt, nhờ các lớp convolution biến dạng	Xuất sắc, transformer giúp nắm bắt chi tiết tốt hơn

Ưu điểm	Nhận diện thời gian thực, hiệu quả trên các thiết bị edge	Linh hoạt cho các cấu trúc phức tạp và biến đổi	Độ chính xác cao, ít lỗi nhận diện sai
Nhược điểm	Khó nhận diện các vật thể nhỏ và chồng chéo	Chậm hơn YOLO, không phù hợp với ứng dụng cần tốc độ	Cần nhiều tài nguyên, tốc độ chậm
Ứng dụng	Ứng dụng thời gian thực: giám sát, xe tự hành, drone	Môi trường phức tạp: công nghiệp, y tế	Ứng dụng yêu cầu độ chính xác cao: lái xe tự động, giám sát từ xa
Yêu cầu phần cứng	Vừa phải (phù hợp với GPU và một số CPU cao cấp)	Trung bình đến cao	Cao (cần GPU/TPU để xử lý hiệu quả)
Yêu cầu dữ liệu huấn luyện	Trung bình	Trung bình đến cao	Cao (yêu cầu nhiều dữ liệu và tuning phức tạp)
Độ phức tạp trong huấn luyện	Đơn giản (YOLO yêu cầu ít tuning hơn)	Trung bình, yêu cầu tuning cho môi trường phức tạp	Rất cao, cần tối ưu hóa nhiều để đạt hiệu suất tốt

2. So sánh kiến trúc CNN-Based và Transformer-Based

Mô hình **CNN-based** (như YOLO, R-CNN) sử dụng các lớp tích chập để trích xuất đặc trưng hình ảnh, thích hợp cho các ứng dụng yêu cầu xử lý thời gian thực nhờ tốc độ cao và hiệu quả tính toán. Các mô hình Single-Stage như YOLO nhanh hơn, nhưng độ chính xác thấp hơn trong các tình huống phức tạp so với các mô hình Two-Stage như Faster R-CNN. Tuy nhiên, CNN bị hạn chế trong việc nắm bắt ngữ cảnh rộng, khiến cho khả năng nhận diện các vật thể nhỏ hoặc bị che khuất bị giảm.

Mô hình **Transformer-Based** (như DETR, DINO-DETR) sử dụng cơ chế tự chú ý (self-attention) để xử lý toàn bộ ngữ cảnh hình ảnh, cho phép phát hiện chính xác hơn, đặc biệt với các vật thể nhỏ hoặc trong môi trường phức tạp. Transformer không cần các anchor boxes hay region proposals, mà dựa vào các truy vấn để phát hiện vật thể, giúp cải thiện khả năng nhận diện. Tuy nhiên, mô hình này đòi hỏi tài nguyên cao và thời gian huấn luyện dài hơn, kém phù hợp cho các ứng dụng thời gian thực nếu không được tối ưu hóa tốt.

3. Kết quả thử nghiệm

Đang tiến hành thực hiện pre-train trên mô hình YOLLO v11 và DINO-DETR ….