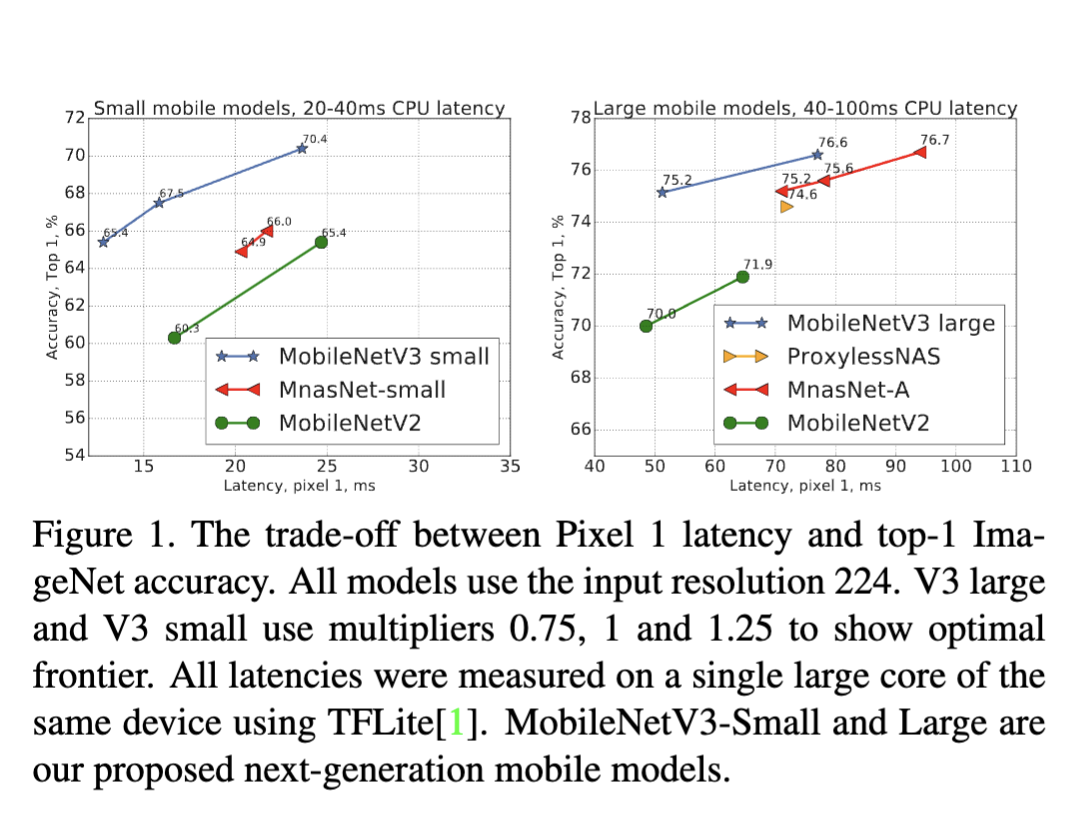
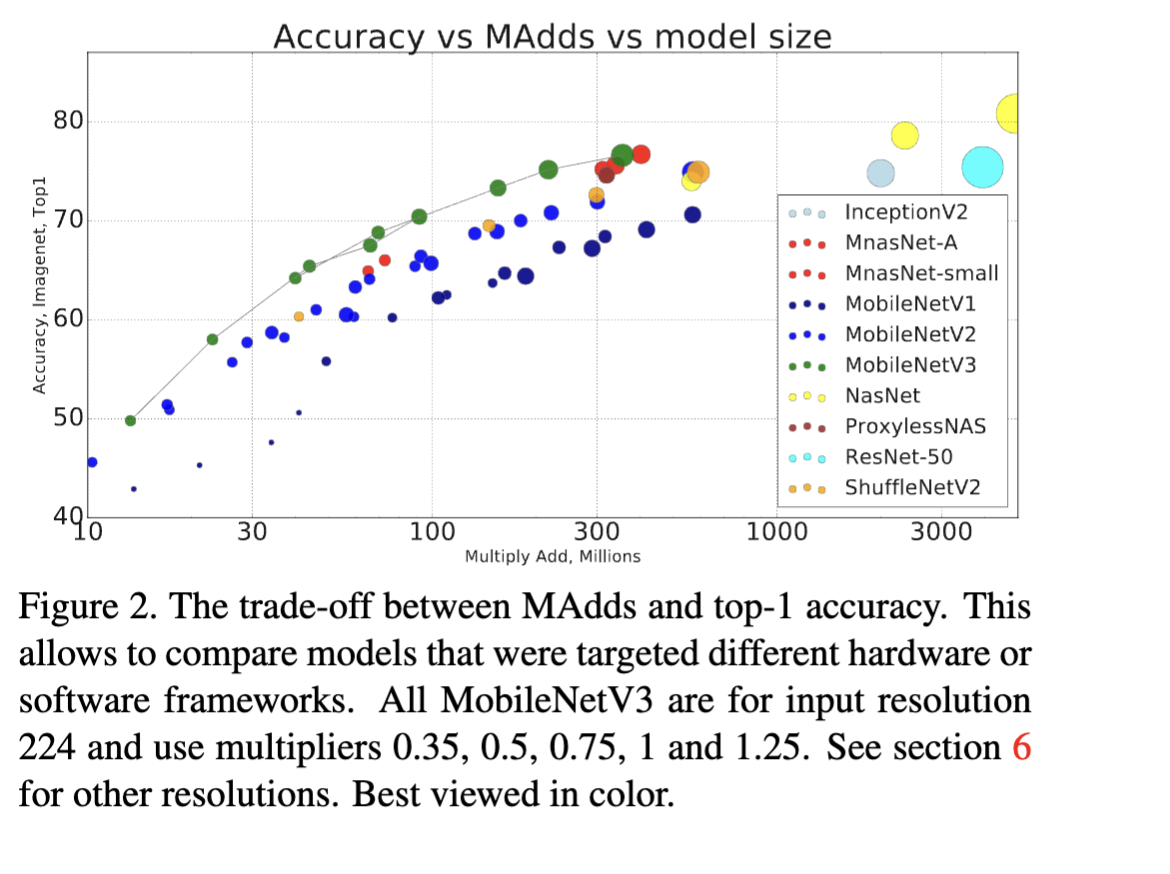
# Tóm tắt bài báo khoa học

📄 MobileNetV3: Searching for MobileNetV3 (Howard et al., 2019)

## 1. Vấn đề và mục tiêu nghiên cứu

Bài báo hướng đến việc thiết kế một mạng nơ-ron hiệu quả cao, có thể hoạt động tốt trên thiết bị di động trong các tác vụ như phân loại, phát hiện và phân đoạn ảnh. Mục tiêu là tối ưu độ chính xác và độ trễ (latency) trên phần cứng thực tế bằng cách kết hợp tìm kiếm kiến trúc mạng (NAS) và cải tiến thiết kế mạng.





## 2. Phương pháp nghiên cứu

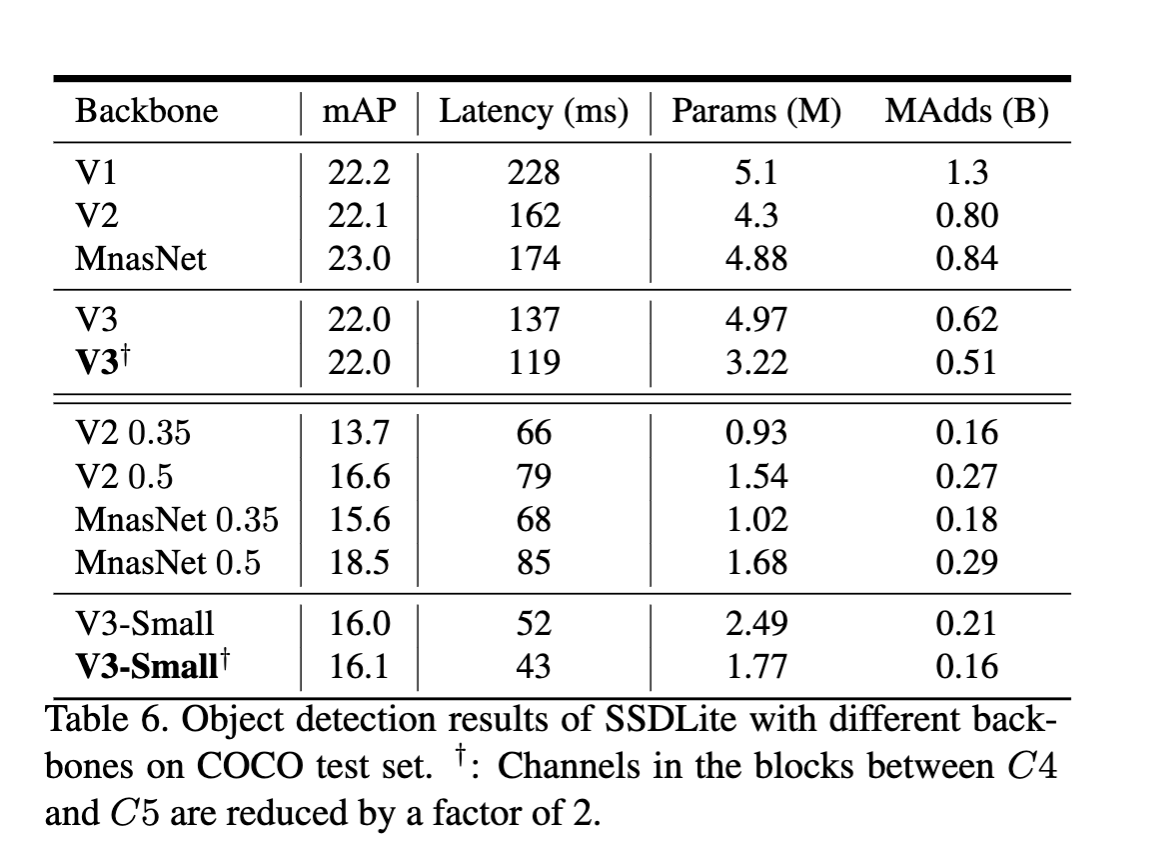
- Sử dụng NAS có ý thức về phần cứng (platform-aware NAS) để tìm ra kiến trúc tổng thể.  
- Áp dụng NetAdapt để tinh chỉnh số lượng bộ lọc cho từng lớp.  
- Giới thiệu h-swish: một dạng đơn giản hóa và hiệu quả của hàm kích hoạt swish.  
- Cải tiến module Squeeze-and-Excite, giảm số lượng tham số và thời gian suy luận.  
- Thiết kế hai mô hình: MobileNetV3-Large (cho thiết bị mạnh), MobileNetV3-Small (cho thiết bị yếu).

## 3. Dữ liệu và thực nghiệm

- Tập dữ liệu: ImageNet (phân loại), COCO (phát hiện), Cityscapes (phân đoạn).  
- Đánh giá trên điện thoại Google Pixel (Pixel 1, 2, 3) về độ chính xác Top-1 và độ trễ.  
- So sánh với các mô hình: MobileNetV2, MnasNet, ProxylessNAS, ShuffleNet, ESPNet, CCC2,...  
- Thực hiện huấn luyện trên TPU với RMSProp, batch size 4096, dropout 0.8, EMA decay 0.9999.

## 4. Kết quả nổi bật

- MobileNetV3-Large đạt Top-1 accuracy 75.2% trên ImageNet, nhanh hơn 20% so với MobileNetV2.  
- MobileNetV3-Small đạt độ chính xác 67.4%, cao hơn MobileNetV2 với độ trễ tương đương.  
- Trong bài toán phát hiện COCO, MobileNetV3-Large nhanh hơn 27% với độ chính xác tương đương.  
- Trong phân đoạn Cityscapes, LR-ASPP nhanh hơn R-ASPP 34% với độ chính xác tương đương.



## 5. Đóng góp chính

- Kết hợp thành công NAS và cải tiến kiến trúc để tối ưu cho thiết bị di động.  
- Đề xuất hàm kích hoạt h-swish hiệu quả và phù hợp với mô hình lượng tử (quantized model).  
- Thiết kế mô hình dễ điều chỉnh và có thể triển khai linh hoạt với tài nguyên khác nhau.  
- Gợi ý sử dụng các thủ thuật giảm kênh, LR-ASPP để tăng tốc mà không giảm độ chính xác.