**MỨC ĐỘ PHẢN HỒI RETINAFACE**

**RetinaFace**

Lý do chọn RetinaFace:

* Độ chính xác rất cao (mAP 96.12% trên WiderFace) và nổi bật với khả năng nhận diện khuôn mặt nhỏ, khuôn mặt bị che khuất, điều kiện ánh sáng yếu, vật cản.
* Đặc biệt mạnh với các trường hợp đặc biệt về khuôn mặt nhờ kiến trúc multi-task (vừa detect vừa landmark).
* Triển khai trung bình về độ phức tạp, có nhiều mã nguồn mở, tài liệu tham khảo và API dễ dùng.
* Yêu cầu tài nguyên vừa phải, cân bằng tốt giữa độ chính xác và hiệu suất (tốt hơn YOLOv5 nếu mục tiêu của bạn là nhận diện nhiều trường hợp khó).
* Có thể mở rộng, cải tiến thêm nếu cần nhận diện trạng thái khuôn mặt đặc biệt (buồn ngủ, mệt mỏi…) nhờ landmark chính xác.

**(Response Latency / Real-time Capability):**

Mức độ phản hồi ở đây được hiểu là tốc độ mà hệ thống có thể xử lý khung hình video và đưa ra kết quả (phát hiện khuôn mặt, tính toán EAR, nhận diện, cảnh báo). Nó thường được đo bằng **Frames Per Second (FPS)** - số khung hình được xử lý mỗi giây, hoặc **Latency (độ trễ)** - thời gian cần thiết để xử lý một khung hình.

**Các yếu tố ảnh hưởng trực tiếp đến mức độ phản hồi:**

1. **Thuật toán Phát hiện Khuôn mặt được chọn:**
   * **Dlib HOG:** Đây là một thuật toán khá nhanh và ổn định trên **CPU**. Mức độ phản hồi có thể đạt từ **15-30+ FPS** trên một CPU tầm trung đến cao, tùy thuộc vào độ phân giải video và số lượng khuôn mặt. Độ trễ tương đối thấp.
   * **RetinaFace (--use-retinaface):**
     + **Trên CPU:** Cực kỳ chậm. Mức độ phản hồi có thể chỉ đạt **dưới 1-5 FPS**, làm cho nó không thực tế cho ứng dụng thời gian thực.
     + **Trên GPU:** Rất nhanh và chính xác. Mức độ phản hồi có thể đạt từ **20-60+ FPS** (hoặc thậm chí cao hơn với GPU cao cấp như RTX 4090, lên tới **40-150+ FPS** tùy phiên bản mô hình RetinaFace và độ phân giải). Đây là lựa chọn tốt nhất cho phản hồi thời gian thực nếu có GPU.
2. **Kích hoạt tính năng Học và Nhận diện Khuôn mặt (--enable-learning và PyTorch):**
   * Tính năng này yêu cầu trích xuất "facial embeddings" và so sánh chúng với cơ sở dữ liệu. Quá trình này **tốn kém tính toán hơn** đáng kể so với chỉ phát hiện khuôn mặt và điểm mốc.
   * **Mức độ phản hồi:** Việc bật tính năng này chắc chắn sẽ **làm giảm FPS và tăng độ trễ**. Mức giảm phụ thuộc vào:
     + Kiến trúc mô hình học sâu được sử dụng để trích xuất embeddings.
     + Kích thước của cơ sở dữ liệu đặc trưng (số lượng người và số lượng embeddings cho mỗi người).
     + Hiệu suất của GPU (nếu có) hoặc CPU.
   * Trên GPU, mức độ phản hồi có thể vẫn ở mức chấp nhận được (ví dụ: **10-30 FPS** tùy thuộc vào các yếu tố trên). Trên CPU, nó có thể trở nên rất chậm (chỉ vài FPS).
3. **Tài nguyên phần cứng (CPU/GPU/RAM):**
   * **CPU:** Ảnh hưởng đến hiệu suất của Dlib HOG và các phần xử lý logic khác. CPU mạnh sẽ cải thiện phản hồi.
   * **GPU:** Quyết định khả năng sử dụng RetinaFace và hiệu suất của tính năng học/nhận diện. **GPU là yếu tố then chốt để đạt được mức độ phản hồi thời gian thực cao** cho các cấu hình nâng cao.
   * **RAM:** Ảnh hưởng đến khả năng tải mô hình lớn và xử lý dữ liệu.
4. **Độ phân giải video đầu vào:**
   * Video độ phân giải cao hơn (Full HD, 4K) sẽ yêu cầu nhiều tính toán hơn và làm giảm FPS, tăng độ trễ. Video độ phân giải thấp hơn (SD, VGA) sẽ cho phản hồi nhanh hơn.
5. **Số lượng khuôn mặt trong khung hình:**
   * Xử lý nhiều khuôn mặt cùng lúc sẽ tăng tải tính toán, đặc biệt nếu tính năng học/nhận diện được bật cho mỗi khuôn mặt.

**Đánh giá mức độ phản hồi theo kịch bản:**

* **Kịch bản 1: Cấu hình cơ bản (Dlib HOG, không học/nhận diện) trên CPU:**
  + **Mức độ phản hồi:** Tốt, phù hợp cho các ứng dụng cơ bản. Có thể đạt **20-30+ FPS** trên CPU tầm trung. Độ trễ thấp.
  + **Đánh giá:** Đáp ứng yêu cầu "thời gian thực" cho đa số trường hợp sử dụng cơ bản.
* **Kịch bản 2: Cấu hình nâng cao (RetinaFace, không học/nhận diện) trên GPU:**
  + **Mức độ phản hồi:** Rất tốt, lý tưởng cho các ứng dụng yêu cầu độ chính xác cao và tốc độ. Có thể đạt **30-60+ FPS** trên GPU tầm trung đến cao. Độ trễ rất thấp.
  + **Đánh giá:** Đạt mức "thời gian thực" cao cấp, phù hợp cho môi trường đòi hỏi độ tin cậy và phản hồi nhanh.
* **Kịch bản 3: Cấu hình đầy đủ (RetinaFace, có học/nhận diện) trên GPU:**
  + **Mức độ phản hồi:** Tốt, nhưng có thể thấp hơn Kịch bản 2. Có thể đạt **10-30 FPS** tùy thuộc vào kích thước mô hình học và cơ sở dữ liệu.
  + **Đánh giá:** Vẫn có thể hoạt động ở chế độ "gần thời gian thực" hoặc "thời gian thực đủ dùng" tùy ứng dụng, nhưng độ trễ sẽ tăng lên so với chỉ phát hiện.
* **Kịch bản 4: Bất kỳ cấu hình nào trên phần cứng yếu (ví dụ: Raspberry Pi, CPU cũ không có GPU):**
  + **Mức độ phản hồi:** Thấp. Có thể chỉ đạt **vài FPS (dưới 10 FPS)**, gây ra độ trễ đáng kể và trải nghiệm không mượt mà.
  + A screenshot of a computer

    AI-generated content may be incorrect.**Đánh giá:** Không phù hợp cho các ứng dụng yêu cầu phản hồi thời gian thực nghiêm ngặt.