**Chiến lược kết hợp**

**Chúng ta có thể tổ chức quá trình so sánh theo các bước sau:**

**Bước 1️: Kiểm tra màu sắc (Histogram Matching) – Lọc nhanh**

Mục đích: Loại bỏ các ảnh không có tông màu giống nhau, giúp giảm số lượng ảnh cần xử lý.

Cách làm:

Tính histogram màu của ảnh truy vấn và ảnh trong tập kết quả.

Dùng khoảng cách Bhattacharyya hoặc Chi-Square để so sánh.

Ngưỡng lọc: Giữ lại top 30-50% ảnh có histogram gần giống nhất ngưỡng khoảng 80%.

✅ Lợi ích: Lọc nhanh, giảm số lượng ảnh cần so sánh chi tiết.

**Bước 2️: Kiểm tra hình dạng & cấu trúc (SSIM)**

Mục đích: Loại bỏ các ảnh có cấu trúc tổng thể khác biệt.

Cách làm:

Chuyển ảnh về grayscale.

Tính toán SSIM giữa ảnh truy vấn và ảnh còn lại sau bước 1.

Ngưỡng lọc: Giữ lại ảnh có SSIM ≥ 0.8.

✅ Lợi ích: Giữ lại các ảnh có bố cục tương tự, giảm ảnh gây nhiễu.

**Bước 3️: So khớp đặc trưng cục bộ (SIFT / ORB)**

Mục đích: Tìm kiếm các điểm đặc trưng khớp nhau giữa ảnh.

(Chỉ chạy SIFT nếu ảnh có nhiều chi tiết > 50k pixel)

Cách làm:

Trích xuất các điểm đặc trưng SIFT hoặc ORB.

Sử dụng FLANN Matcher hoặc BFMatcher để so khớp.

Ngưỡng lọc: Giữ lại ảnh có >50 keypoints trùng khớp.

✅ Lợi ích: Xác định chi tiết giống nhau ngay cả khi ảnh bị xoay, biến đổi kích thước.

**Bước 4️: So sánh deep features (DINOv2 + Cosine Similarity)**

Mục đích: So sánh nội dung ảnh ở mức độ sâu.

Chỉ gọi DINO nếu các phương pháp khác không đủ tốt

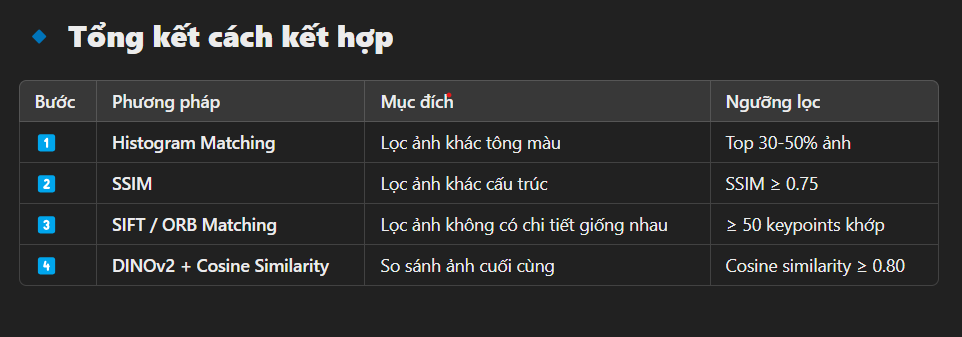
Cách làm:

Trích xuất vector đặc trưng bằng DINOv2.

Tính cosine similarity giữa ảnh truy vấn và ảnh còn lại từ bước 3.

Ngưỡng lọc: Giữ lại ảnh có cosine similarity ≥ 0.80.

✅ Lợi ích: Xác định chính xác ảnh có nội dung tương đồng.



**🔹 Tại sao nên kết hợp cả 4 phương pháp?**

✅ Lọc nhanh ngay từ đầu bằng Histogram → Tăng tốc độ.

✅ Loại bỏ ảnh sai lệch ngay từ đầu bằng SSIM.

✅ Đảm bảo độ chính xác cao bằng SIFT/ORB.

✅ Dùng AI để so sánh sâu bằng DINOv2, giúp tăng độ chính xác cuối cùng

**Quy trình chi tiết:**

1. **Tìm kiếm hình ảnh liên quan trên Google**:
   * Truy cập vào Google, tìm kiếm hình ảnh liên quan và lấy kết quả đầu tiên.
   * Truy cập vào trang kết quả khớp chính xác và trích xuất toàn bộ nội dung bài báo từ các trang này và trả về kết quả.
   * Đóng trình duyệt google tại đây.
2. **Nếu không có bài báo nào**:
   * Chuyển qua trang hình ảnh trùng khớp của Google và tải về 10 hình ảnh đầu tiên vào một folder tạm.
   * Khi hoàn thành việc tải 10 hình ảnh đầu tiên về thì đóng trình duyệt tại đây.
3. **Lọc hình ảnh ra khỏi folder tạm**:
   * **Bước 1 (Histogram Matching)**: Kiểm tra màu sắc của các ảnh, loại bỏ các ảnh không có tông màu giống nhau.
   * **Bước 2 (SSIM)**: So sánh cấu trúc của các hình ảnh và giữ lại những hình ảnh có chỉ số SSIM ≥ 0.80.
4. **Nếu không có best\_image. Sau bước 1 và bước 2 trong folder còn lại bao nhiêu hình ảnh thì tiến hành trích xuất đặc trưng để sử dụng ở bước 3 và 4**:
   * **Bước 3 (SIFT)**: So khớp các đặc trưng cục bộ giữa các hình ảnh nếu độ phân giải của hình ảnh >= 50k pixel.
   * **Bước 4 (DINOv2 + Cosine Similarity)**: So sánh các đặc trưng sâu của hình ảnh và giữ lại ảnh có cosine similarity ≥ 0.80.
5. **Tìm kiếm lại hình ảnh đã lọc tốt nhất trên Google**:
   * Sau khi lọc được best\_image, thực hiện tìm kiếm lại best\_image đó trên Google và lấy kết quả bài báo liên quan.

.