Nhận dạng khuôn mặt dựa trên một phần với Vision Máy biến áp

Zhonglin Sun
zhonglin.sun@qmul.ac.uk

Georgios Tzimiropoulos
g.tzimiropoulos@qmul.ac.uk

Khoa Kỹ thuật Điện tử và
Khoa học máy tính
Đại học Queen Mary London
Luân Đôn, Vương quốc Anh

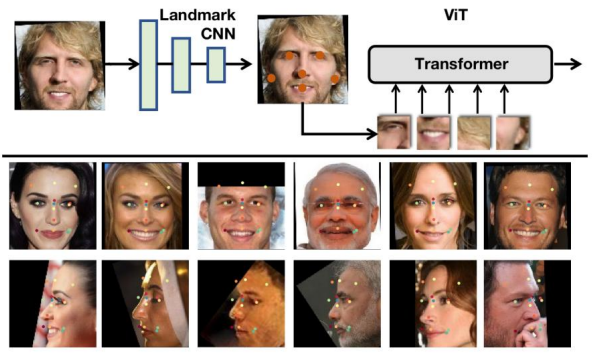
Tóm tắt

Các phương pháp toàn diện sử dụng CNN và tổn thất dựa trên biên độ đã thống trị nghiên cứu về nhận dạng khuôn mặt. Trong công trình này, chúng tôi tách khỏi bối cảnh này theo hai cách: (a) chúng tôi sử dụng Vision Transformer làm kiến trúc để đào tạo đường cơ sở rất mạnh cho nhận dạng khuôn mặt, được gọi đơn giản là fViT, vốn đã vượt qua hầu hết các phương pháp nhận dạng khuôn mặt tiên tiến. (b) Thứ hai, chúng tôi tận dụng đặc tính vốn có của Transformer để xử lý thông tin (mã thông báo trực quan) được trích xuất từ các lưới không đều để thiết kế một đường ống cho nhận dạng khuôn mặt gợi nhớ đến các phương pháp nhận dạng khuôn mặt dựa trên một phần. Đường ống của chúng tôi, được gọi là một phần fViT, chỉ đơn giản bao gồm một mạng nhẹ để dự đoán tọa độ của các điểm mốc trên khuôn mặt theo sau là Vision Transformer hoạt động trên các bản vá được trích xuất từ các điểm mốc được dự đoán và được đào tạo từ đầu đến cuối mà không có sự giám sát của điểm mốc. Bằng cách học cách trích xuất các bản vá phân biệt, Transformer dựa trên một phần của chúng tôi tiếp tục tăng cường độ chính xác của đường cơ sở Vision Transformer của chúng tôi, đạt được độ chính xác tiên tiến trên một số điểm chuẩn nhận dạng khuôn mặt.

1 Giới thiệu

Nhận dạng khuôn mặt (FR) là một vấn đề quan trọng trong thị giác máy tính với nhiều ứng dụng như kiểm soát biên giới và giám sát. Với sự ra đời của Học sâu, đường ống thực tế cho FR trong những năm gần đây bao gồm (a) xương sống CNN (Mạng nơ-ron tích chập), xử lý hình ảnh khuôn mặt một cách toàn diện để tính toán nhúng đặc điểm khuôn mặt được sử dụng để tính điểm tương đồng và (b) hàm mất mát thích hợp cho việc học nhúng phân biệt. Trong khi phần lớn công trình gần đây về FR tập trung vào (b), tức là thiết kế các hàm mất mát hiệu quả hơn [8, 13, 33, 37, 48, 58, 61], công trình này chủ yếu tập trung vào (a) tức là thiết kế các kiến trúc mới để trích xuất đặc điểm khuôn mặt.

Động lực đầu tiên của công trình của chúng tôi là Vision Transformer [16] mới được giới thiệu gần đây, đang ngày càng phổ biến trong Computer Vision với các kết quả gần đây được báo cáo là rất cạnh tranh với các kết quả do xương sống CNN tạo ra [38, 65]. Do đó, đóng góp đầu tiên của chúng tôi là khám phá xem người ta có thể đi xa đến đâu với ViT vani để nhận dạng khuôn mặt bằng cách sử dụng mất mát vani của [58]. Chúng tôi chỉ ra rằng một xương sống như vậy với tối ưu hóa siêu tham số thích hợp đã đạt được kết quả tiên tiến nhất cho nhận dạng khuôn mặt. Động lực thứ hai cho công trình của chúng tôi là ViT, trái ngược với CNN, thực sự có thể hoạt động trên các bản vá được trích xuất



Hình 1: Minh họa về ViT dựa trên bộ phận của chúng tôi để nhận dạng khuôn mặt. Một hình ảnh khuôn mặt được xử lý bởi CNN mốc nhẹ tạo ra một tập hợp các mốc khuôn mặt. Các mốc được sử dụng để lấy mẫu các bộ phận khuôn mặt từ hình ảnh đầu vào, sau đó được sử dụng làm đầu vào cho ViT để trích xuất và nhận dạng đặc điểm. Toàn bộ hệ thống được đào tạo từ đầu đến cuối mà không có sự giám sát mốc. Ví dụ về các mốc được CNN mốc phát hiện được hiển thị.

từ các lưới không đều và không yêu cầu lưới lấy mẫu cách đều được sử dụng cho các phép tích chập. Vì khuôn mặt người là một vật thể có cấu trúc bao gồm các bộ phận (ví dụ: mắt, mũi, môi) và được lấy cảm hứng từ công trình quan trọng về nhận dạng khuôn mặt dựa trên bộ phận trước khi học sâu [5], trong bài báo này, chúng tôi đề xuất áp dụng ViT trên các mảng biểu diễn các bộ phận trên khuôn mặt. Cụ thể, đóng góp thứ hai của chúng tôi là một đường ống dựa trên bộ phận mới được đề xuất để nhận dạng khuôn mặt sâu, trong đó các điểm mốc được học một cách phân biệt trước tiên được dự đoán thông qua một CNN điểm mốc nhẹ, các mảng được trích xuất xung quanh chúng và sau đó đưa vào ViT. Đáng chú ý là toàn bộ hệ thống, được gọi là phần fViT, có thể được đào tạo từ đầu đến cuối mà không cần giám sát điểm mốc. Hình 1 cho thấy tổng quan về đường ống được đề xuất.

Tóm lại, những đóng góp của chúng tôi là:

- Chúng tôi đào tạo một ViT thông thường để nhận dạng khuôn mặt một cách phù hợp bằng cách sử dụng một mất mát thông thường, mà chúng tôi gọi là fViT, và chứng minh rằng fViT tạo ra kết quả tiên tiến nhất trên một số điểm chuẩn nhận dạng khuôn mặt phổ biến.

- Chúng tôi tận dụng kiến trúc Transformer để đề xuất một đường ống mới cho nhận dạng khuôn mặt, được đặt tên là phần fViT, trong đó các bản vá được học phân biệt trước tiên được trích xuất và sau đó đưa vào ViT để nhận dạng, về cơ bản là xây dựng một ViT dựa trên một phần để nhận dạng khuôn mặt. Đáng chú ý là CNN mốc được sử dụng để dự đoán các mốc được đào tạo từ đầu đến cuối với ViT mà không có giám sát mốc.

- Chúng tôi chứng minh rằng phần fViT của chúng tôi vượt qua fViT cơ sở mạnh mẽ của chúng tôi, thiết lập một trạng thái mới của nghệ thuật trên một số tập dữ liệu nhận dạng khuôn mặt. Hơn nữa, chúng tôi loại bỏ một số thành phần trong đường ống của mình để minh họa tác động của chúng đối với độ chính xác của nhận dạng khuôn mặt.

- Chúng tôi chứng minh rằng CNN mang tính bước ngoặt, một phần trong đường ống của chúng tôi, có hiệu quả đối với nhiệm vụ phụ là khám phá các điểm mốc không có giám sát.

2 Công trình liên quan

Đánh giá chi tiết về các bài báo về nhận dạng khuôn mặt nằm ngoài phạm vi, ở đây chúng tôi tập trung vào các vấn đề về tổn thất, phươ ng pháp nhận biết vùng và Vision Transformers có liên quan nhiều hơ n đến công việc của chúng tôi.

Chức năng mất mát: Một số bài báo [8, 13, 33, 37, 48, 58, 61] đã tập trung vào các tính năng học tập có thể tách biệt và phân biệt thông qua việc sử dụng chức năng mất mát thích hợp. Trong khi khả năng tách biệt có thể đạt được với mất mát softmax, việc học các đặc điểm phân biệt khó hơ n vì trong lô nhỏ, quá trình đào tạo không thể thấy được sự phân phối đặc điểm toàn cục [61]. Để đạt được mục đích này, FaceNet [48] sử dụng bộ ba để học trực tiếp phép ánh xạ vào không gian Euclidean nhỏ gọn sao cho các đặc điểm khuôn mặt từ cùng một danh tính càng gần nhau càng tốt trong khi các đặc điểm từ các danh tính khác nhau càng xa nhau càng tốt.

Để tránh vấn đề lựa chọn bộ ba, Center loss [61] giảm thiểu khoảng cách giữa các đặc điểm sâu đã học cho mỗi khuôn mặt và các tâm lớp tươ ng ứng của chúng để đạt được sự tập trung trong lớp. Nhận thấy rằng các ranh giới giữa các lớp không được tách biệt tốt trong Softmax Loss, L-softmax [36] xem xét công thức kết hợp của softmax cross-entropy loss và lớp tuyến tính, phạt khoảng cách của ranh giới lớp, dẫn đến các đặc điểm phân biệt hơ n. Sau đó, CosFace[58] áp dụng chuẩn hóa không chỉ trên các trọng số mà còn trên nhúng đặc điểm và đề xuất thêm biên độ trên cos(θ) trong đó θ là góc giữa trọng số tuyến tính và nhúng. ArcFace[9] định nghĩa thêm biên độ trên góc θ thay vì cos(θ). VPL[13] chú ý đến việc học nguyên mẫu của từng lớp bằng cách xem xét phân phối của các lớp trên không gian đặc điểm và đề xuất thay đổi nguyên mẫu tĩnh bằng cách đưa các đặc điểm đã ghi nhớ vào để xấp xỉ biến thể nguyên mẫu.

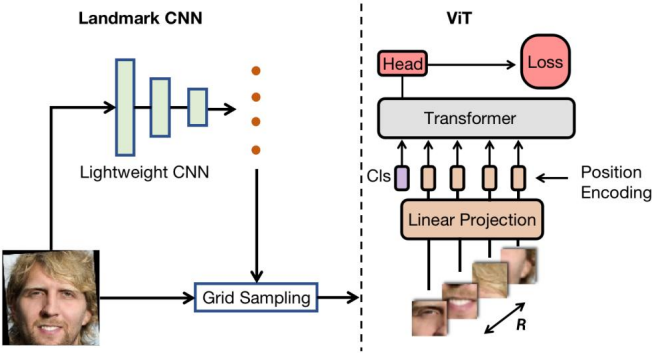
Gần đây, Sphereface2 [62] đề xuất tiến hành phân loại nhị phân để nhận dạng và một số nguyên tắc chung cũng được tóm tắt trong công trình về cách thiết kế tổn thất tốt.

Phươ ng pháp nhận biết vùng: Mặc dù CNN cung cấp xươ ng sống tiêu chuẩn cho nhận dạng khuôn mặt dựa trên thông tin toàn cục, nhưng chúng bỏ qua thực tế rằng khuôn mặt là một đối tượng có cấu trúc với các phần có thể được sử dụng để học hiệu quả hơ n các đặc điểm khuôn mặt. Ví dụ, công trình quan trọng của [5], là công trình tiên tiến nhất trước khi học sâu ra đời, cho thấy rằng việc trích xuất một số lượng rất lớn các đặc điểm đa tỷ lệ xung quanh 5 điểm mốc được xác định trước (ví dụ: mắt, mũi, miệng) có thể rất hiệu quả để nhận dạng khuôn mặt. Để giải quyết các đặc điểm cục bộ thông qua các giải pháp dựa trên học sâu, TUA [35] đã đề xuất tích hợp các đặc điểm khuôn mặt cục bộ và toàn cục từ các CNN rời rạc khác nhau thông qua các GPU khác nhau, để tổng hợp hoạt động nối đặc điểm được sử dụng. FAN-Face [67] đã khám phá cách các đặc điểm từ mạng định vị điểm mốc trên khuôn mặt được đào tạo trước có thể được sử dụng để nâng cao độ chính xác của nhận dạng khuôn mặt, tuy nhiên mạng định vị và nhận dạng điểm mốc không được đào tạo chung. Hơ n nữa, [15, 25, 26] đều đưa ra các phươ ng pháp để trích xuất các đặc điểm liên quan đến mốc trong quá trình đào tạo CNN, tuy nhiên, chúng vẫn yêu cầu các mốc được xác định trước. Để tránh giám sát mốc rõ ràng, Comparator Networks [66] đề xuất một đường ống thực hiện chú ý đến nhiều vùng cục bộ phân biệt bản địa (mốc) và sử dụng chúng để so sánh các mô tả cục bộ giữa các cặp khuôn mặt. Cuối cùng, HPD [59] tận dụng tối đa cơ chế chú ý để dự đoán mặt nạ chú ý cho các đặc điểm cục bộ.

Phần fViT của chúng tôi lấy cảm hứng từ [5, 66] nhưng hoạt động theo cách hoàn toàn khác. Đầu tiên, các điểm mốc được học bằng cách trực tiếp dự đoán tọa độ x,y của chúng bằng cách sử dụng một mạng rất nhẹ (tức là mobilenetV3 [20]). Sau đó, các bản vá có tâm tại các điểm mốc được dự đoán được lấy mẫu và đưa vào Transformer [16, 57] để nhận dạng khuôn mặt. Đáng chú ý là chúng tôi tận dụng kiến trúc Transformer để cung cấp đầu vào là một tập hợp các bản vá được lấy mẫu tại các vị trí không gian bất thường, khác với các phươ ng pháp nhận dạng khuôn mặt tiêu chuẩn dựa trên CNN sử dụng lưới hình ảnh thông thường (cần thiết để xác định tích chập) nhưng cũng khác với ViT [16] cũng

sử dụng lưới thông thường để xử lý hình ảnh đầu vào. Hơ n nữa, hệ thống của chúng tôi được đào tạo theo cách đầu cuối mà không cần giám sát mốc.

Vision Transformer: Transformer lần đầu tiên được giới thiệu trong Natural Language Pro-processing cho dịch máy và các tác vụ NLP khác [57]. Nó bao gồm các lớp Self-attention và Feed-Forward. Vision Transformer (ViT) được giới thiệu trong [16] và kể từ đó đã được chứng minh là cung cấp độ chính xác cạnh tranh cho CNN [65]. Đào tạo ViT khó hơ n so với CNN [55, 56]. Một số phươ ng pháp đã được đề xuất để tạo điều kiện thuận lợi cho việc đào tạo ViT [6, 17, 38, 52, 56, 60, 64, 65, 69, 70, 70]. Trong công trình này, chúng tôi loại bỏ phươ ng pháp trước đây sử dụng ViT để nhận dạng khuôn mặt [77] trong đó các bản vá chồng chéo thông thường được trích xuất từ khuôn mặt, thay vào đó chúng tôi áp dụng xươ ng sống ViT tiêu chuẩn [16] với các cải tiến đào tạo của [52]. Điều này đã cung cấp cho chúng tôi một đường cơ sở rất mạnh vượt qua hầu hết các phươ ng pháp tiên tiến hiện có để nhận dạng khuôn mặt trên tập dữ liệu MS1M [18] . Tiếp theo, chúng tôi đi xa hơ n [16] và các công trình tiếp theo [6, 17, 38, 52, 56, 60, 64, 65, 69, 70, 70] bằng cách áp dụng bộ biến áp, lần đầu tiên theo hiểu biết của chúng tôi trên một tập hợp các bản vá được trích xuất từ các lưới không đều do mạng nhẹ cung cấp, được đào tạo từ đầu đến cuối để cung cấp các điểm mốc phân biệt mà không cần giám sát rõ ràng.



Hình 2: Cấu trúc tổng thể của phần fViT mà chúng tôi đề xuất: Một CNN nhẹ được sử dụng để dự đoán một tập hợp các điểm mốc trên khuôn mặt. Sau đó, lấy mẫu lưới có thể phân biệt được áp dụng để trích xuất các phần khuôn mặt phân biệt được sau đó được sử dụng làm đầu vào cho ViT để trích xuất và nhận dạng đặc điểm. Các nút màu vàng biểu thị tọa độ điểm mốc trên khuôn mặt đã hồi quy được trích xuất

3 Phươ ng pháp

Trong Phần 3.1, trước tiên chúng tôi mô tả đường cơ sở mạnh của mình, được gọi là fViT, thu được bằng cách đào tạo ViT với mất mát CosFace. Sau đó, trong Phần 3.2, chúng tôi giới thiệu ViT dựa trên một phần được đề xuất của chúng tôi để nhận dạng khuôn mặt, được gọi là một phần fViT.

3.1 fViT: ViT để nhận dạng khuôn mặt

Chúng tôi được cung cấp một hình ảnh khuôn mặt $X \in \mathbb{R}^{H\times W\times C}$ (C = 3). Theo ViT [16], hình ảnh được chia thành các mảng không chồng lấn $R = P \times P$ sau đó được ánh xạ thành các mã thông báo trực quan bằng cách sử dụng lớp nhúng tuyến tính $P \in \mathbb{R}^{2\times d}$ E R · Để bảo tồn thông tin không gian, nhúng vị trí

P 2×d
ps R cũng được học và được thêm vào các mã thông báo trực quan ban đầu. Sau đó, chuỗi mã thông báo được xử lý bởi các lớp L Transformer.

l Mã thông báo trực quan tại lớp l và vị trí không gian s là z 0,...,P đến chuỗi mã thông s d R , l = 0,...,L 1, s = d R cls báo [14]. 2 l 1. Ngoài các mã thông báo trực quan R, một mã thông báo phân loại z được thêm vào trước

Lớp Biến áp thứ l xử lý các mã thông báo trực quan Z R (P 2+1)×d của lớp trước đó bằng cách sử dụng một loạt Tự chú ý nhiều đầu (MSA), Lớp

Chuẩn hóa (LN) và MLP (R 4dd R R) các lớp như sau:

$$\text{Có} = \text{MSA}(\text{LN}(Z^{l-1})) + Z^{l-1}, \quad (1)$$

$$Z = \text{MLP}(\text{LN}(Y)) + \text{Có}. \quad (2)$$

Một đầu Tự chú ý (SA) duy nhất được đưa ra bởi:

$$s = \sum_{s=0}^{P^2} \sigma\{(q_s^l \cdot k_s)/\sqrt{dh}\} v_s, s = 0,\dots,P^2 1, \quad (3)$$

trong đó σ(.) = Softmax(.), q s ,k s R dh là các vectơ truy vấn, khóa và giá trị sử đưa ra từ sự s dụng ma trận nhúng Wq,Wk,Wv R d×dh , dh là hệ số tỷ lệ trong tự chú ý z . Cuối cùng, đầu ra của h đầu được nối lại và chiếu bằng cách sử dụng ma trận nhúng Wh R hdh×d .

Mã thông báo phân loại z cls L được đào tạo để nhận dạng khuôn mặt bằng cách sử dụng mất mát CosFace [58]:

$$\text{Mất mát} = \frac{1}{N} \log \frac{\text{và}^{b(\cos(\theta yi,i) m)}}{\text{và}^{b(\cos(\theta yi,i) m)} \text{bcos}(\theta j,i) + j=yi e}, \quad (4)$$

trong đó N là số mẫu trong một lô, z = ||z $\frac{z_{\text{lớp}}^{L}}{L}$, zi là mẫu thứ i và yi là cls|| giá trị thực tế tương ứng, W = cột $\frac{W}{||W||}$ là ma trận trọng số của lớp tuyến tính cuối cùng, Wj là thứ j được chuẩn hóa (lớp) của ma trận trọng số, cos(θyi ,i) = WT và b được cố định là ||z cls||. L zi m là biên độ

Chúng tôi thấy rằng fViT, tương tự như ViT, dễ bị quá khớp. Do đó, để đạt được độ chính xác cao, chúng tôi đã sử dụng kết hợp các phương pháp đào tạo bao gồm chính quy hóa độ sâu ngẫu nhiên [30], thay đổi kích thước & cắt ngẫu nhiên, RandAugment [7], Cutout và cuối cùng là Mixup [71]. Chi tiết về lựa chọn của những phương pháp này được đưa ra trong tài liệu bổ sung 2.1.

3.2 Phần fViT

ViT như được mô tả bởi các Phương trình 1 & 2 hoạt động trên một chuỗi mã thông báo trực quan không cần phải tính toán trên lưới đồng nhất. Lấy cảm hứng từ công trình về FR dựa trên bộ phận [5], trong phần này chúng tôi mô tả cách áp dụng ViT trên các mảng biểu diễn các bộ phận trên khuôn mặt.

Cụ thể, chúng tôi sử dụng CNN có trọng số nhẹ để dự đoán một tập hợp các điểm mốc R = P×P:

$$r = \text{CNN}(X), r_i = [x_i, y_i]^T, \text{tôi} = 1,\dots,P^2, \quad (5)$$

trong đó chúng tôi đã sử dụng MobilenetV3 [20] cho CNN của mình.

Sau đó, chúng tôi lấy mẫu một bản vá có tâm tại mỗi tọa độ điểm mốc ri . Để phù hợp với tọa độ phân số, chúng tôi đã sử dụng phương pháp lấy mẫu lưới có thể phân biệt của STN [23] để trích xuất từng bản vá. Sau đó, mỗi bản vá được mã hóa bằng lớp nhúng

E, tạo ra các mã thông báo phần R cùng với mã thông báo lớp được xử lý bởi Bộ biến đổi của Phươ ng trình 1 & 2. Chúng tôi khám phá một số tùy chọn cho mã hóa vị trí được thêm vào mã thông báo phần trong nghiên cứu cắt bỏ ở Phần 4.2.

Toàn bộ đường ống, được gọi là phần fViT, rất đơ n giản và được thể hiện trong Hình 2. Đường ống được đào tạo từ đầu đến cuối mà không có giám sát mốc chỉ bằng cách sử dụng mất mát CosFace của Phươ ng trình 4. Đáng chú ý là mạng hồi quy mốc tạo thành một nút thắt thông tin trước đây được thấy hữu ích trong các phươ ng pháp khám phá mốc không giám sát [24]. Chúng tôi cũng xác nhận phát hiện này trong một nghiên cứu cắt bỏ trong Phần 4.2. Cuối cùng, mặc dù có thể sử dụng các phươ ng pháp hồi quy bản đồ nhiệt với softmax, chúng tôi đã chọn hồi quy tọa độ trực tiếp đơ n giản hơ n.

4 Thí nghiệm

Trong phần này, chúng tôi đánh giá độ chính xác của các bộ biến đổi mặt được đề xuất trên một số tập dữ liệu nổi tiếng và so sánh chúng với độ chính xác của các phươ ng pháp tiên tiến mới được đề xuất gần đây.

4.1 Chi tiết triển khai

Để đào tạo và để so sánh công bằng với các phươ ng pháp khác, chúng tôi đã sử dụng phiên bản tinh chỉnh [10] của MS1M [18] (MS1MV3) chứa 93.431 danh tính trừ khi được chỉ định. Chúng tôi cũng cung cấp kết quả đào tạo trên VGGFace2 [3] với 3,1 triệu hình ảnh và 8,6 nghìn danh tính. Hình ảnh khuôn mặt có độ phân giải 112 × 112 và được căn chỉnh (do [9] cung cấp) Chúng tôi đã thử nghiệm các mô hình của mình trên LFW [21], CFP-FP [50], AgeDB-30 [43], IJB-B [63], IJB-C [41] và MegaFace [27] để tiến hành đánh giá hiệu suất nhận dạng. Đối với LFW, CFP-FP và AgeDB-30, chúng tôi sử dụng độ chính xác xác minh 1:1 (%). Chúng tôi báo cáo kết quả TAR@FAR=1e-4 trên IJB-B và IJB-C. Đối với Megaface, Megaface/id đề cập đến độ chính xác nhận dạng hạng 1 (%) trên 1M bộ phân tán, và Megaface/ver đề cập đến độ chính xác xác minh TAR@FAR=1e-6. Để đào tạo Trans-former, chúng tôi đã chọn sử dụng một lượng lớn dữ liệu tăng cường so với thiết lập FR ban đầu được sử dụng trong ResNet, vui lòng tham khảo tài liệu bổ sung Phần 2.1.1 và 2.1.2 để biết thêm chi tiết về siêu tham số, tăng cường, cấu trúc mô hình và chi tiết đào tạo.

4.2 Nghiên cứu cắt bỏ

Chúng tôi đã tiến hành một số nghiên cứu để làm nổi bật tác động của các lựa chọn thiết kế khác nhau cho các máy biến áp khuôn mặt của chúng tôi. Các nghiên cứu cắt bỏ của chúng tôi chủ yếu được thực hiện trên miếng vá số R = 49 vì tốc độ đào tạo hiệu quả của nó. Chúng tôi cũng đính kèm phần cải thiện về tăng cường dữ liệu, mức độ chồng chéo và Hiệu ứng của các CNN mốc khác nhau trong phần tài liệu bổ sung 2.2.

Hiệu ứng của số lượng bản vá và các mô hình fViT khác nhau: Thí nghiệm đầu tiên của chúng tôi tập trung vào cách số lượng bản vá (hoặc tươ ng đươ ng là số lượng điểm mốc R cho phần fViT) tác động đến độ chính xác của các Bộ biến đổi khuôn mặt được đề xuất. Số lượng bản vá được chọn là 16, 49 và 196 với FLOP lần lượt là 1,17G, 3,3G và 12,64G và cả hai mô hình fViT-B và fViT-S đều được thử nghiệm, như minh họa trong Bảng 1. Lưu ý rằng khi số lượng bản vá tăng lên, kích thước bản vá K giảm xuống; cụ thể đối với 196 điểm mốc, kích thước bản vá tươ ng ứng là 8 và đối với 16 điểm mốc, kích thước bản vá là 28, đảm bảo rằng đối với trường hợp số lượng điểm mốc nhỏ, toàn bộ ảnh khuôn mặt vẫn được phân tích. fViT-B có đặc điểm mờ với 768 và

MLP mờ với 2048 trong khi fViT-S có 512 và MLP mờ với 2560. Trong cả hai trường hợp, số của đầu là 11. Kết quả được thể hiện trong Bảng 1.

Một số kết luận thú vị có thể được rút ra từ thí nghiệm này: (1) Nhiều bản vá hơn (các điểm mốc) dẫn đến dự đoán chính xác hơn, như mong đợi. (2) Khi số lượng các bản vá (điểm mốc) rất lớn (tức là 196) thì phần fViT vượt trội hơn fViT một chút. (3) Khi số lượng bản vá/điểm mốc giảm, khoảng cách này tăng lên cụ thể đối với CFP-FP và AgeDB. Điều này quan trọng vì các mô hình xử lý ít mã thông báo hơn đáng kể nhẹ. Ví dụ, mô hình 49 landmark nhanh hơn 4 lần so với mô hình 196 landmark.

Miếng vá xương sống số 196		Mô hình LFW CFP-FP AgeDB IJB-C			
fViT-B	196	phần fViT 99,83 99,21 98,29 97,29			
	49	fViT 99,85 99,01 98,13 97,21			
	49	phần fViT 99,80 98,78 97,85 96,37			
	16	fViT 97,56 96,38 98,00 phần			
	16	fViT 99,80 97,30 97,22 94,90			
		fViT 99,78 96,87 96,46 94,85			
fViT-S	196	phần fViT 99,83 fViT	99,09	98,18	96,58
	196	99,83 phần fViT	98,90	97,90	96,50
	49	99,80 fViT 99,80	98,7	97,81	96,33
	49		98,0	97,31	96,05
	16	phần fViT 99,71 fViT	97,25	97,06	94,21
	16	99,71	96,95	96,25	94,19

Bảng 1: Tác động của số lượng bản vá và các mô hình fViT khác nhau đến độ chính xác của FR.

Tác động của các mã hóa vị trí khác nhau Ở đây, chúng tôi khám phá chức năng của mã hóa vị trí mã hóa trong phần fViT-B R = 49 điểm mốc của chúng tôi. Chúng tôi kiểm tra 3 loại mã hóa vị trí: (a) những cái có thể đào tạo được như trong fViT gốc [16], (b) cosin [57] và (c) dựa trên tọa độ. Đối với dựa trên tọa độ, chúng tôi đã sử dụng một lớp tuyến tính để nhúng mỗi điểm mốc ri vào Rd và sau đó thêm vectơ này đến mã thông báo trực quan tương ứng. Kết quả được hiển thị trong Bảng 2 (phần trên cùng). Như Có thể thấy phương pháp có thể đào tạo và phương pháp dựa trên tọa độ đạt được độ chính xác tốt nhất.

Cuộc thí nghiệm	Nội dung	LFW CFP-FP AgeDB IJB-C			
Mã hóa vị trí	Có thể huấn luyện được	99,80 98,78 97,85 96,37			
	Cô sin	99,80 98,65 98,03 96,08			
	Điều phối	99,80 98,71 97,66 96,29			
Nút thắt thông tin	với IB	99,80	98,78	97,85	96,37
	không có IB	99,76	97,73	97,31	96,05
Điểm mốc không được giám sát	Vani fViT 99,78 phần fViT (MobilenetV3)		98,00	97,56	96,30
	99,80 phần fViT (FAN (Frozen)) 99,36 phần fViT		98,78	97,85	96,37
	(MobilenetV3 (Frozen)) 99,81		95,31	96,11	93,96
			98,72	97,66	96,35

Bảng 2: Kết quả của các nghiên cứu cắt bỏ khác nhau: (a) Phần trên cùng: tác động của các vị trí khác nhau mã hóa. (b) Phần giữa: tác động của nút thắt thông tin. (c) Phần cuối: tác động của khám phá mốc không giám sát. Tất cả các thí nghiệm đều có phần fViT-B với R = 49.

Tác động của nút thắt thông tin: Chúng tôi đã thử nghiệm cung cấp cho bộ phận fViT như nhập tính năng của lớp áp chót từ CNN mang tính bước ngoặt, về cơ bản là đưa vào các tính năng từ CNN đến fViT và vi phạm nút thắt thông tin của đường ống của chúng tôi trong Phần 3.2. Cụ thể, tính năng của lớp áp chót CNN đã được nối với (đào tạo

có thể) mã hóa vị trí và sau đó chiếu tới Rd . Kết quả được thể hiện trong Bảng 2 (giữa phần). Như đã quan sát, việc vi phạm nút thắt thông tin sẽ làm giảm độ chính xác.

Hiệu ứng của việc khám phá điểm mốc không giám sát: Vì các phương pháp định vị điểm mốc trên khuôn mặt có giám sát được sử dụng rộng rãi trong tài liệu, chúng tôi so sánh phần fViT của mình với một mô hình sử dụng các điểm mốc được cung cấp bởi công nghệ định vị điểm mốc khuôn mặt hiện đại, cụ thể là FAN [2]. Chúng tôi đóng băng phần CNN mang tính bước ngoặt từ phần fViT được đào tạo bài bản để đào tạo một ViT mới, được đặt tên là như một phần fViT(mobilenet (Frozen)). Kết quả được hiển thị trong Bảng 2 (phần dưới cùng). Vì nó có thể được quan sát, sử dụng FAN (Các tham số được đào tạo trước và đóng băng) để cung cấp các điểm mốc đầu vào đến fViT làm giảm hiệu suất xuống mức dưới mức tối ưu. Cách này sử dụng trực tiếp các bản vá lỗi của các điểm mốc được cung cấp bởi một mạng lưới mốc được giám sát chính xác dẫn đến kết quả tệ hơn so với việc đào tạo một vanilla fViT. Với mạng lưới mốc R=49 được đào tạo trước và chỉ đào tạo phần fViT, chúng tôi đạt được sự cải thiện đáng kể so với mạng FAN. Chúng ta có thể kết luận rằng đối với trực tiếp khi sử dụng các bản vá điểm mốc trên nhiệm vụ FR, FAN không thể cung cấp các điểm mốc thích hợp.

Phương pháp	LFW	CFP-FP	AgeDB	IJB-B	IJB-C	MegaFace/id	MegaFace/ver	
CosFace[58]	99,81	94,80	96,37	97,91	98,12	97,33	94,25	96,03 98,35 GroupFace[28] 98,85 97,91
94,93	96,26	98,74	CircleLoss[53]	99,73	93,95	98,50	COL[4]	99,83 94,61 98,60
CurricularFace[22]	99,80	94,8	96,1	98,71	98,37	97,62	96,20	98,48
Face[67]	99,85	94,97	96,38	98,70	BroadFace[29]	99,85	94,97	96,38 98,70 ArcFace-challenge[12] 98,79
99,85	96,81	VPL[13]	99,83	95,56	96,76	ALN[73]	93,13	95,27 VirFace[32] 88,90 90,54 98,22 98,73
MagFace[42]	94,51	95,97	94,74	96,09	SCL[33]	98,37	98,32	98,78 -
				98,80		98,31		98,64
				98,63		98,38		98,36
				98,63		98,38		98,95
				99,06		98,48		98,95
				99,11		98,60		- - -
		-		96,53		97,25		98,80 98,97
		99,56		97,15		-		- -
		99,83		98,46		96,15		- -
		99,80		98,59		98,26		81,40 97,15
Biến đổi khuôn mặt [77]	95,96	99,83		96,19		97,82	-	- -
fViT-B, của chúng tôi		99,85		99,01		98,13	95,97	97,21 98,69 98,91
Phần fViT-B, của chúng tôi		99,83		99,21		98,29	96,11	97,29 98,96 98,78

Bảng 3: So sánh với tình trạng hiện tại trên nhiều tập dữ liệu. fViT cơ bản của chúng tôi và phần fViT đạt được kết quả tiên tiến nhất trên hầu hết các tập dữ liệu.

4.3 So sánh với công nghệ tiên tiến

Chúng tôi đã chọn phần fViT-B và fViT của mình với kích thước miếng vá 8 và R=196 để so sánh với gần đây đề xuất các phương pháp FR tiên tiến nhất. CNN mang tính bước ngoặt được sử dụng là MobilenetV3.

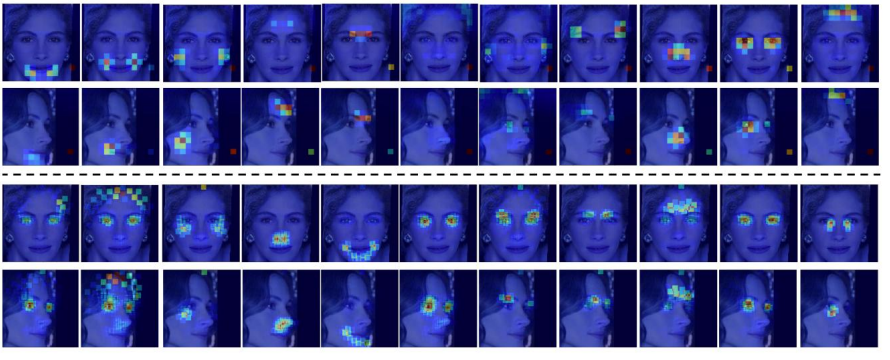
Kết quả định lượng: Chúng tôi báo cáo kết quả của các mô hình được đào tạo trên MS1MV3 và được thử nghiệm trên nhiều chuẩn mực khác nhau. Kết quả được thể hiện trong Bảng 3. Như đã quan sát, trên LFW là bão hòa, các phương pháp chúng tôi đề xuất đạt được độ chính xác cao nhất cùng với một số phương pháp khác. Trên tập dữ liệu nhạy cảm với tư thế CFP-FP, phần fViT của chúng tôi đã đạt được độ chính xác là 99,21%, vượt qua các phương pháp tiên tiến khác của VPL [13] và Arcface-challenge[12]. Tương tự kết quả được quan sát thấy đối với các chuẩn mực IJB-B và IJB-C: không chỉ phần fViT của chúng tôi vượt trội hơn các phương pháp tiên tiến khác với biên độ đáng kể (97,29 TAR trên IJB-C, 96,11 TAR trên IJB-B), nhưng ngay cả fViT cơ sở của chúng tôi cũng là phương pháp tốt thứ hai (97,21 TAR trên IJB-C và 95,97 TAR trên IJB-B). Kết quả tương tự thu được trên đánh giá MegaFace, nơi chúng tôi

phần fViT là hiệu suất cao nhất cùng với một vài phương pháp khác. Ngoại lệ duy nhất là AgeDB-30, trong đó phần fViT của chúng tôi đạt được 98,29%. Chúng tôi cần đề cập rằng hàm mất mát được sử dụng là CosFace [58] được chọn vì tính đơn giản và ổn định của nó. Có thể rằng sử dụng các hàm mất mát tiên tiến hơn để đào tạo, bao gồm VPL [13], ArcFace [9] và Sphereface2 [62]. Chúng tôi cũng đã tiến hành các thí nghiệm trên tập dữ liệu VGGFace2 bằng cách sử dụng tương tự

	LFW	AgeDB-30	IJB-B	IJB-C	MegaFace/Id	MegaFace/Ver	
Mạng so sánh [66]	-	-		85,0	88,5		
FAN-Face [67]	-	-		91,1	93,5	89,41	-
Mặt cầu [37]	99,55	92,88		91,96	88,61	71,53	85.02
Khuôn mặt Cos [58]	99,51	92,98		90,98	89,11	71,65	85,45
Mặt vòng cung [9]	99,47	91,97		91,60	88,56	73,65	87,77
Mất mát vòng tròn [53]	99,48	92,90		90,83	91,31	93,25	71,32	84,34
SphereFace2 [62] fViT,	99,50	93,68				74,38	89,19
Phần của chúng	99,44	93,52		88,13	90,26		71,11	85,04
tôi fViT, Của chúng tôi	99,56	93,92		88,98	91,03		71,63	85,91

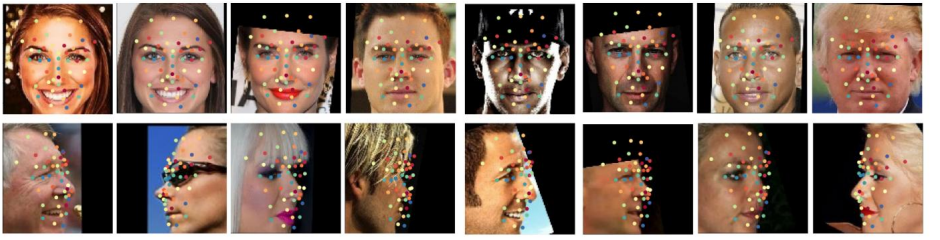
Bảng 4: So sánh với kết quả hiện tại trên VGGFace2.

tham số với Resnet64 trong SphereFace [37] để hiển thị kết quả của phần fViT của chúng tôi trong Bảng 4. Mặc dù thêm một lượng lớn dữ liệu tăng cường, fViT cơ sở của chúng tôi hoạt động kém hơn kết quả được cung cấp bởi Resnet64 tương tự như Face Transformer khi đào tạo trên một tập dữ liệu quy mô nhỏ như CASIA-webface [68]. Phần fViT của chúng tôi cũng đạt được hiệu quả tốt hơn kết quả hơn fViT ban đầu khi đào tạo trên MS1M, trong khi nó vẫn tệ hơn một chút so với Resnet64 với các tổn thất nâng cao (ví dụ: ArcFace [9]). Công việc tương lai của chúng tôi sẽ điều tra cách thức phương pháp này hoạt động trên các chuẩn mực quy mô lớn khác như Glink360 [1].



Hình 3: Hình ảnh hóa bản đồ chú ý. Hàng đầu tiên và hàng thứ hai hiển thị 11 chú ý bản đồ được tạo ra bởi 11 đầu của fViT-B cơ sở; Hàng thứ ba và thứ tư hiển thị 11 bản đồ chú ý được tạo ra bởi 11 đầu của phần fViT-B với R = 196 điểm mốc.

Kết quả định tính: Đầu tiên chúng tôi so sánh các bản đồ chú ý được tạo ra bởi 11 người đứng đầu đường cơ sở fViT và phần fViT trong Hình 3. Chúng tôi quan sát thấy rằng đối với cả hai phương pháp, đầu đạt được sự tương ứng tốt giữa các tư thế khi mỗi đầu bắn vào các khu vực tương ứng ở cả phía trước và hình ảnh hồ sơ. Sau đó, nhìn kỹ hơn sẽ thấy rằng sự chú ý của người đứng đầu thứ 6 và thứ 7 (thứ 6 và cột thứ 7 của Hình 3) của fViT cơ sở (hàng thứ 1 và hàng thứ 2) không tập trung vào cụ thể các bộ phận trên khuôn mặt. Hơn nữa, đối với fViT cơ bản chỉ có một đầu tập trung vào mắt. Điều này hoàn toàn trái ngược với phần fViT, nơi có nhiều người đứng đầu tập trung vào vùng mắt được biết đến là bộ phận trên khuôn mặt phân biệt rõ nhất đối với FR [31,



Hình 4: Hình ảnh hóa các điểm mốc đã học được từ phần fViT-B của chúng tôi với R = 49. Các điểm mốc có cùng màu trong các hình ảnh khác nhau ở các tư thế đã được học ở một mức độ nào đó.

45, 49, 59, 66, 72]. Hình 4 cho thấy 49 điểm mốc được học bởi phần fViT của chúng tôi. Như đã hiển thị, sự tươ ng ứng điểm mốc giữa các tư thế đã được học ở một mức độ tốt. Bên cạnh kết quả FR, CNN điểm mốc của chúng tôi có thể hữu ích để cung cấp các điểm mốc trên khuôn mặt được học mà không cần giám sát điểm mốc. Có thể xem giải thích chi tiết trong phần tài liệu bổ sung Mục 2.3

5 Kết luận

Chúng tôi đề xuất Face Transformers làm kiến trúc cho nhận dạng khuôn mặt có độ chính xác cao. Chúng tôi đã mô tả hai mô hình: (a) fViT, đường cơ sở mạnh của chúng tôi được đào tạo phù hợp trên MS1M. (b) phần fViT, chúng tôi đã tận dụng thuộc tính của Transformer để xử lý các mã thông báo trực quan được trích xuất từ các lưới không đều để đề xuất một Face Transformer dựa trên một phần được đào tạo từ đầu đến cuối để thực hiện định vị mốc và nhận dạng khuôn mặt mà không cần giám sát mốc rõ ràng. Đường ống của chúng tôi cực kỳ đơ n giản bao gồm một CNN nhẹ để hồi quy tọa độ trực tiếp theo sau là một ViT hoạt động trên các bản vá được trích xuất từ các mốc dự đoán. Cả hai mô hình, và đặc biệt là phần fViT của chúng tôi, đều đạt được độ chính xác tiên tiến hoặc gần tiên tiến trên một số điểm chuẩn nhận dạng khuôn mặt.

Sự thừa nhận

Zhonglin Sun được hỗ trợ bởi Hội đồng học bổng Trung Quốc (CSC).

Tài liệu tham khảo

[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Fc một phần: Đào tạo 10 triệu danh tính trên một máy. Trong Kỷ yếu của Hội nghị Quốc tế IEEE/CVF về Thị giác Máy tính, trang 1445–1449, 2021.

[2] Adrian Bulat và Georgios Tzimiropoulos. Chúng ta còn cách giải quyết vấn đề căn chỉnh khuôn mặt 2d & 3d bao xa? (và một tập dữ liệu gồm 230.000 điểm mốc khuôn mặt 3d). Trong Biên bản báo cáo Hội nghị quốc tế IEEE về thị giác máy tính, trang 1021–1030, 2017.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi và Andrew Zisserman. Vggface2: Một tập dữ liệu để nhận dạng khuôn mặt theo tư thế và độ tuổi. Trong FG, 2018.

[4] Jie Chang, Zhonghao Lan, Changmao Cheng và Yichen Wei. Học dữ liệu không chắc chắn trong nhận dạng khuôn mặt. Trong Biên bản báo cáo Hội nghị IEEE/CVF về máy tính Tầm nhìn và Nhận dạng Mẫu, trang 5710–5719, 2020.

[5] Dong Chen, Xudong Cao, Fang Wen và Jian Sun. Phước lành của tính đa chiều: Tính năng đa chiều và nén hiệu quả của nó để xác minh khuôn mặt. Trong Biên bản của hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, trang 3025–3032, 2013.

[6] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei và Qi Tian. Visformer: Máy biến áp thân thiện với thị giác. Bản in trước của arXiv arXiv:2104.12533, 2021.

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens và Quoc V Le. Randaugment: Tăng cường dữ liệu tự động thực tế với không gian tìm kiếm được thu hẹp. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Thị giác máy tính và Hội thảo nhận dạng mẫu, trang 702–703, 2020.

[8] Đặng Kiến Khang, Giả Quốc, Niannan Xue và Stefanos Zafeiriou. Arcface: Phụ gia mất biên góc cho nhận dạng khuôn mặt sâu. Trong CVPR, 2019.

[9] Đặng Kiến Khang, Giả Quốc, Niannan Xue và Stefanos Zafeiriou. Arcface: Phụ gia mất biên góc cho nhận dạng khuôn mặt sâu. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 4690–4699, 2019.

[10] Đặng Kiến Khang, Giả Quốc, Trương Đức Bình, Đặng Á Phong, Lộ Tương Cư, và Tống Thực. Thử thách nhận dạng khuôn mặt nhẹ. Trong Biên bản báo cáo của IEEE/CVF International Hội nghị về Hội thảo Thị giác máy tính, trang 0–0, 2019.

[11] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong và Stefanos Zafeiriou. Sub-center arcface: Tăng cường nhận dạng khuôn mặt bằng khuôn mặt web nhiễu quy mô lớn. Ở châu Âu Hội nghị về Thị giác máy tính, trang 741–757. Springer, 2020.

[12] Đặng Kiến Khang, Giả Quốc, Tương An, Zheng Zhu, và Stefanos Zafeiriou. Mặt nạ thách thức nhận dạng: Báo cáo theo dõi insightface. Trong Biên bản báo cáo của IEEE/CVF Hội nghị quốc tế về thị giác máy tính, trang 1437–1444, 2021.

[13] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas và Stefanos Zafeiriou. Học nguyên mẫu biến thể để nhận dạng khuôn mặt sâu. Trong Biên bản báo cáo của IEEE/CVF Hội nghị về Thị giác máy tính và Nhận dạng mẫu, trang 11906–11915, 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee và Kristina Toutanova. Bert: Đào tạo trước các bộ biến đổi song hướng sâu để hiểu ngôn ngữ. Bản in trước arXiv arXiv:1810.04805, 2018.

[15] Changxing Ding và Dacheng Tao. Mạng nơ-ron tích chập nhánh thân cây cho nhận dạng khuôn mặt dựa trên video. Giao dịch IEEE về phân tích mẫu và trí tuệ máy móc, 40(4):1002–1014, 2017.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiao-hua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit và Neil Houlsby. Một hình ảnh có giá trị 16x16 từ: Máy biến áp để nhận dạng hình ảnh ở quy mô lớn. Trong Hội nghị quốc tế

về Biểu diễn học tập, 2021. URL https://openreview.net/forum? id=YicbFdNTTy.

[17] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou và Matthijs Douze. Levit: một máy biến đổi tầm nhìn trong trang phục của hội đồng để suy luận nhanh hơ n. bản in trước arXiv arXiv:2104.01136, 2021.

[18] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He và Jianfeng Gao. Ms-celeb-1m: Một tập dữ liệu và chuẩn mực cho nhận dạng khuôn mặt quy mô lớn. Trong hội nghị châu Âu về thị giác máy tính, trang 87–102. Springer, 2016.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren và Jian Sun. Học dư sâu để nhận dạng hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 770–778, 2016.

[20] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Tìm kiếm mobilenetv3. Trong Biên bản Hội nghị quốc tế IEEE/CVF về Tầm nhìn máy tính, trang 1314–1324, 2019.

[21] Gary B Huang, Marwan Mattar, Tamara Berg và Eric Learned-Miller. Khuôn mặt được gắn nhãn trong tự nhiên: Cơ sở dữ liệu để nghiên cứu nhận dạng khuôn mặt trong môi trường không bị hạn chế. Trong Hội thảo về khuôn mặt trong 'Hình ảnh thực tế': phát hiện, căn chỉnh và nhận dạng, 2008.

[22] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li và Feiyue Huang. Curricularface: mất mát học tập chươ ng trình giảng dạy thích ứng cho nhận dạng khuôn mặt sâu. Trong biên bản báo cáo hội nghị IEEE/CVF về thị giác máy tính và nhận dạng mẫu, trang 5901–5910, 2020.

[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Mạng lưới biến áp không gian. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 28:2017–2025, 2015.

[24] Tomas Jakab, Ankush Gupta, Hakan Bilen và Andrea Vedaldi. Học không giám sát các điểm mốc của đối tượng thông qua việc tạo ảnh có điều kiện. Trong Biên bản báo cáo Hội nghị quốc tế lần thứ 32 về Hệ thống xử lý thông tin thần kinh, trang 4020–4031, 2018.

[25] Bong-Nam Kang, Yonghyun Kim và Daijin Kim. Mạng quan hệ từng cặp để nhận dạng khuôn mặt. Trong Biên bản báo cáo Hội nghị châu Âu về tầm nhìn máy tính (ECCV), trang 628–645, 2018.

[26] Bong-Nam Kang, Yonghyun Kim, Bongjin Jun và Daijin Kim. Mạng quan hệ cặp tính năng phân cấp để nhận dạng khuôn mặt. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Thị giác máy tính và Hội thảo nhận dạng mẫu, trang 0–0, 2019.

[27] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller và Evan Brossard. Chuẩn mực khuôn mặt lớn: 1 triệu khuôn mặt để nhận dạng ở quy mô lớn. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 4873–4882, 2016.

[28] Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, và Jongju Shin. Groupface: Học các nhóm tiềm ẩn và xây dựng các biểu diễn dựa trên nhóm để nhận dạng khuôn mặt. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 5621–5630, 2020.

[29] Yonghyun Kim, Wonpyo Park, và Jongju Shin. Broadface: Nhìn vào hàng chục nghìn người cùng một lúc để nhận dạng khuôn mặt. Trong Hội nghị Châu Âu về Thị giác Máy tính, trang 536–552. Springer, 2020.

[30] Gustav Larsson, Michael Maire và Gregory Shakhnarovich. Fractalnet: Mạng nơ -ron siêu sâu không có phần dư. Bản in trước arXiv arXiv:1605.07648, 2016.

[31] Susan J Lederman, Roberta L Klatzky và Ryo Kitada. Xử lý khuôn mặt xúc giác và mối quan hệ của nó với thị giác. Trong Nhận thức đối tượng đa giác quan trong não linh trưởng, trang 273–300. Springer, 2010.

[32] Pengyu Li, Biao Wang và Lei Zhang. Lớp kết nối hoàn toàn ảo: Đào tạo tập dữ liệu nhận dạng khuôn mặt quy mô lớn với tài nguyên tính toán hạn chế. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 13315–13324, 2021.

[33] Shen Li, Jianqing Xu, Xiaqing Xu, Pengchen Shen, Shaoxin Li và Bryan Hooi. Học tập sự tự tin hình cầu để nhận dạng khuôn mặt. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 15629–15637, 2021.

[34] Weijian Li, Haofu Liao, Shun Miao, Le Lu và Jiebo Luo. Học không giám sát các điểm mốc trên khuôn mặt dựa trên sự nhất quán giữa các chủ thể. Trong Hội nghị quốc tế lần thứ 25 về nhận dạng mẫu (ICPR) năm 2020, trang 4077–4082. IEEE, 2021.

[35] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei và Chang Huang. Đạt được độ chính xác tối đa khi nhắm mục tiêu: Nhận dạng khuôn mặt thông qua tính năng nhúng sâu. bản in trước arXiv arXiv:1506.07310, 2015.

[36] Weiyang Liu, Yandong Wen, Zhiding Yu và Meng Yang. Tổn thất softmax biên độ lớn cho mạng nơ -ron tích chập. Trong ICML, tập 2, trang 7, 2016.

[37] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, và Le Song. Sphereface: Nhúng siêu cầu sâu để nhận dạng khuôn mặt. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 212–220, 2017.

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin và Baining Guo. Bộ chuyển đổi Swin: Bộ chuyển đổi tầm nhìn phân cấp sử dụng cửa sổ dịch chuyển. Trong Biên bản báo cáo Hội nghị quốc tế về tầm nhìn máy tính IEEE/CVF (ICCV), trang 10012–10022, tháng 10 năm 2021.

[39] Ilya Loshchilov và Frank Hutter. Chính quy hóa suy giảm trọng số tách biệt. Bản in trước arXiv arXiv:1711.05101, 2017.

[40] Dimitrios Mallis, Enrique Sanchez, Matthew Bell và Georgios Tzimiropoulos. Học không giám sát các điểm mốc của đối tượng thông qua sự tương ứng tự đào tạo. Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, 33, 2020.

[41] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Bộ dữ liệu khuôn mặt và giao thức. Trong Hội nghị quốc tế về sinh trắc học (ICB) năm 2018, trang 158–165. IEEE, 2018.

[42] Qiang Meng, Shichao Zhao, Zhida Huang và Feng Zhou. Magface: Một biểu diễn phổ quát cho nhận dạng khuôn mặt và đánh giá chất lượng. Trong Biên bản báo cáo của IEEE/CVF Hội nghị về Thị giác máy tính và Nhận dạng mẫu, trang 14225–14234, 2021.

[43] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia và Stefanos Zafeiriou. Agedb: lần đầu tiên được thu thập thủ công, ngoài tự nhiên cơ sở dữ liệu tuổi. Trong Biên bản báo cáo của Hội nghị IEEE về Tầm nhìn máy tính và Mẫu Hội thảo công nhận, trang 51–59, 2017.

[44] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga và Adam Lerer. Tự động phân biệt trong pytorch. 2017.

[45] Rajeev Ranjan, Vishal M Patel và Rama Chellappa. Hyperface: Một nhiệm vụ đa dạng sâu sắc khuôn khổ học tập để phát hiện khuôn mặt, xác định vị trí mốc, ước tính tư thế và nhận dạng giới tính. Giao dịch IEEE về phân tích mẫu và trí thông minh máy móc, 41 (1):121–135, 2017.

[46] Mihir Sahasrabudhe, Zhixin Shu, Edward Bartrum, Riza Alp Guler, Dimitris Sama-ras và Iasonas Kokkinos. Nâng cao bộ mã hóa tự động: Học không giám sát mô hình 3 chiều có thể biến đổi hoàn toàn bằng cách sử dụng cấu trúc không cứng sâu từ chuyển động. Trong Biên bản báo cáo của Hội nghị quốc tế IEEE/CVF về Hội thảo thị giác máy tính, trang 0–0, 2019.

[47] Enrique Sanchez và Georgios Tzimiropoulos. Phát hiện mốc đối tượng thông qua sự thích nghi không giám sát. Bản in trước arXiv arXiv:1910.09469, 2019.

[48] Florian Schroff, Dmitry Kalenichenko và James Philbin. Facenet: Một nhúng thống nhất cho nhận dạng khuôn mặt và phân cụm. Trong Biên bản hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, trang 815–823, 2015.

[49] Philippe G Schyns, Lizann Bonnar và Frédéric Gosselin. Hãy cho tôi xem các tính năng! hiểu biết về sự nhận dạng từ việc sử dụng thông tin trực quan. Khoa học tâm lý, 13(5):402–409, 2002.

[50] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chel-lappa và David W Jacobs. Xác minh khuôn mặt từ chính diện đến nghiêng trong tự nhiên. Năm 2016 Hội nghị mùa đông IEEE về ứng dụng của thị giác máy tính (WACV), trang 1–9. IEEE, 2016.

[51] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Para-gios và Iasonas Kokkinos. Bộ mã hóa tự động biến dạng: Gỡ rối không giám sát hình dạng và vẻ ngoài. Trong Biên bản báo cáo của hội nghị châu Âu về máy tính thị lực (ECCV), trang 650–665, 2018.

[52] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszko-reit và Lucas Beyer. Làm thế nào để đào tạo vit của bạn? dữ liệu, tăng cường và chính quy hóa trong máy biến áp thị giác. bản in trước arXiv arXiv:2106.10270, 2021.

[53] Tôn Yifan, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, và Yichen Wei. Mất mát vòng tròn: Một quan điểm thống nhất về tối ưu hóa độ tương đồng của cặp. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Mẫu Nhận dạng, trang 6398–6407, 2020.

[54] James Thewlis, Hakan Bilen và Andrea Vedaldi. Học không giám sát các điểm mốc đối tượng bằng cách nhúng không gian nhân tử. Trong Biên bản báo cáo hội nghị quốc tế IEEE về thị giác máy tính, trang 5916–5925, 2017.

[55] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablay-rolles, và Hervé Jégou. Đào tạo dữ liệu biến đổi hình ảnh hiệu quả và chắt lọc thông qua sự chú ý. Trong Hội nghị quốc tế về Học máy, trang 10347–10357. PMLR, 2021.

[56] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve và Hervé Jégou. Đi sâu hơ n với bộ chuyển đổi hình ảnh. bản in trước arXiv arXiv:2103.17239, 2021.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser và Illia Polosukhin. Sự chú ý là tất cả những gì bạn cần. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 5998–6008, 2017.

[58] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li và Wei Liu. Cosface: Mất cosin biên độ lớn cho nhận dạng khuôn mặt sâu. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 5265–5274, 2018.

[59] Qiangchang Wang, Tianyi Wu, He Zheng và Guodong Guo. Mạng lưới chú ý đa dạng theo kim tự tháp phân cấp để nhận dạng khuôn mặt. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 8326–8335, 2020.

[60] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo và Ling Shao. Bộ chuyển đổi tầm nhìn kim tự tháp: Một xươ ng sống linh hoạt cho dự đoán dày đặc mà không cần tích chập. Bản in trước arXiv arXiv:2102.12122, 2021.

[61] Yandong Wen, Kaipeng Zhang, Zhifeng Li và Yu Qiao. Một phươ ng pháp học đặc điểm phân biệt để nhận dạng khuôn mặt sâu. Trong hội nghị châu Âu về thị giác máy tính, trang 499–515. Springer, 2016.

[62] Yandong Wen, Weiyang Liu, Adrian Weller, Bhiksha Raj và Rita Singh. Sphereface2: Phân loại nhị phân là tất cả những gì bạn cần để nhận dạng khuôn mặt sâu. Bản in trước arXiv arXiv:2108.01513, 2021.

[63] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Bộ dữ liệu khuôn mặt Iarpa janus benchmark-b. Trong biên bản hội nghị IEEE về hội thảo về thị giác máy tính và nhận dạng mẫu, trang 90–98, 2017.

[64] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan và Lei Zhang. Cvt: Giới thiệu về phép tích chập trong bộ chuyển đổi thị giác. Bản in trước arXiv arXiv:2103.15808, 2021.

[65] Tete Xiao, Piotr Dollar, Mannat Singh, Eric Mintun, Trevor Darrell và Ross Gir-shick. Các phép tích chập ban đầu giúp máy biến áp nhìn tốt hơ n. Trong A. Beygelzimer, Y. Dauphin, P. Liang và J. Wortman Vaughan, biên tập viên, Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id=Lpfh1Bpqfk.

[66] Weidi Xie, Li Shen và Andrew Zisserman. Mạng so sánh. Trong Biên bản báo cáo của Hội nghị châu Âu về tầm nhìn máy tính (ECCV), trang 782–797, 2018.

[67] Jing Yang, Adrian Bulat, và Georgios Tzimiropoulos. Fan-face: một cải tiến trực giao đơn giản cho nhận dạng khuôn mặt sâu. Trong Biên bản báo cáo Hội nghị về Trí tuệ nhân tạo AAAI, tập 34, trang 12621–12628, 2020.

[68] Dong Yi, Zhen Lei, Shengcai Liao và Stan Z Li. Học cách biểu diễn khuôn mặt từ scratch. bản in trước arXiv arXiv:1411.7923, 2014.

[69] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Chu, Fengwei Yu và Wei Wu. Kết hợp các thiết kế tích chập xếp hạng vào các máy biến áp trực quan. bản in trước arXiv arXiv:2103.11816, 2021.

[70] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng và Shuichen Yan. Tokens-to-token vit: Đào tạo người biến đổi tầm nhìn từ đầu trên imagenet. bản in trước arXiv arXiv:2101.11986, 2021.

[71] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, và David Lopez-Paz. hỗn hợp: Vượt ra ngoài việc giảm thiểu rủi ro theo kinh nghiệm. Bản in trước của arXiv arXiv:1710.09412, 2017.

[72] Lei Zhang, Meng Yang, Xiangchu Feng, Yi Ma và David Zhang. Phân loại dựa trên sự biểu hiện-phản đối hợp tác để nhận dạng khuôn mặt. Bản in trước arXiv arXiv:1204.2358, 2012.

[73] Yaobin Zhang, Weihong Deng, Yaoyao Zhong, Jiani Hu, Xian Li, Dongyue Zhao và Dongchao Wen. Làm sạch nhiễu nhãn thích ứng với siêu giám sát để nhận dạng khuôn mặt sâu. Trong Biên bản báo cáo Hội nghị quốc tế về thị giác máy tính IEEE/CVF, trang 15065–15075, 2021.

[74] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He và Honglak Lee. Phát hiện không giám sát các điểm mốc đối tượng dưới dạng biểu diễn cấu trúc. Trong Biên bản báo cáo Hội nghị IEEE về Tầm nhìn máy tính và Nhận dạng mẫu, trang 2694–2703, 2018.

[75] Zhanpeng Zhang, Ping Luo, Chen Change Loy và Xiaoou Tang. Phát hiện điểm mốc trên khuôn mặt bằng cách học đa nhiệm sâu. Trong hội nghị châu Âu về thị giác máy tính, trang 94–108. Springer, 2014.

[76] Zhanpeng Zhang, Ping Luo, Chen Change Loy và Xiaoou Tang. Học cách biểu diễn sâu để căn chỉnh khuôn mặt với các thuộc tính phụ trợ. Giao dịch IEEE về phân tích mẫu và trí thông minh máy móc, 38(5):918–930, 2015.

[77] Yaoyao Zhong và Weihong Deng. Biến đổi khuôn mặt để nhận dạng. Bản in trước arXiv arXiv:2103.14803, 2021.

Giới thiệu

Đây là tài liệu bổ sung cho bài báo Nhận dạng khuôn mặt dựa trên một phần với Tầm nhìn
Máy biến áp. Đầu tiên chúng tôi trình bày lựa chọn chi tiết về việc tăng cường dữ liệu mà chúng tôi đã sử dụng để nâng cao
fViT trong Phần B.1.1. Sau đó, chúng tôi liệt kê các chi tiết mô hình được áp dụng cho fViT của chúng tôi trong Phần B.1.2.
Hiệu ứng của việc tăng cường dữ liệu, tỷ lệ chồng chéo của các điểm mốc và so sánh lựa chọn
của CNN mốc cũng được đưa vào như nghiên cứu cắt bỏ bổ sung trong Phần B.2. Cuối cùng, chúng tôi mô tả mạng mốc đã
học có hiệu quả cho nhiệm vụ phụ của Ứng dụng
khám phá mốc không có giám sát trong Phần B.3.

B Bổ sung vào phần 4: Thí nghiệm

B.1 Chi tiết triển khai

B.1.1 Chi tiết đào tạo

Để đào tạo Transformer, chúng tôi đã chọn sử dụng một lượng lớn dữ liệu tăng cường so với
với thiết lập nhận dạng khuôn mặt ban đầu được sử dụng trong ResNets. Cụ thể, chúng tôi đã sử dụng ngẫu nhiên
chính quy hóa độ sâu với xác suất 0,1 [30], thay đổi kích thước & cắt trong phạm vi [0,9,1,0], RandAug-ment [7]
với độ lớn là 2 và không có các hoạt động solarize và invert, Mixup [71] với
alpha=0,5 và xác suất là 0,2, Cắt bỏ với giá trị 0,1 và giảm trọng lượng 1e-1 cho ViT
xương sống và 5e-2 cho CNN Landmark. Chúng tôi đã áp dụng AdamW [39] và tốc độ suy giảm học tập cosin theo sau là
khởi động trong 5 kỷ nguyên, trong khi chúng tôi đã đào tạo tổng cộng trong 34 kỷ nguyên. Tất cả
mạng lưới được đào tạo từ đầu.

Mô hình	Kích thước ẩn	Tham số	FLOPS
phần fViT-B	768	66M	12,64G
fViT-B	768	63M	12,58G
Resnet-100	-	65M	12.10G
phần fViT-S	512	46 triệu	8,96G
fViT-S	512	43 triệu	8,90G
Resnet-50	-	43,59 triệu	6,33G

Bảng 5: Kích thước mạng và FLOPS cho fViT và Part fViT và Resnet của chúng tôi

B.1.2 Chi tiết mô hình

Để so sánh công bằng với Resnet [9] được sử dụng làm xương sống trong hầu hết các phương pháp gần đây, chúng tôi
đã xây dựng fViT của chúng tôi để có kích thước mô hình và FLOPS tương tự với Resnet-100.

Cấu hình cơ sở cho fViT, được gọi là fViT-B, có 12 lớp, 11 đầu chú ý và d = 768.
Chúng tôi cũng đã xây dựng một fViT-S. Các mô hình của chúng tôi và Resnet-100 được so sánh trong Bảng 5. Như có thể
quan sát, fViT-B của chúng tôi có kích thước mô hình và FLOPS tương tự với Resnet-100. Điểm mốc của chúng tôi
mạng là MobilenetV3 [20] trừ khi được chỉ định khác. Tất cả các mô hình được triển khai trong
PyTorch [44].

B.2 Nghiên cứu cắt bỏ bổ sung

B.2.1 Hiệu ứng của các phép tăng cường dữ liệu khác nhau

Ở đây chúng tôi trình bày hiệu quả của việc lựa chọn các phần tăng cường khác nhau được đề xuất trong [52] bắt đầu từ filp ngẫu nhiên. Kết quả có thể được tìm thấy trong 6, chúng ta có thể quan sát thấy rằng với nhiều dữ liệu hơn Nếu thêm phần tăng cường, kết quả sẽ chính xác hơn.

Exp	Flip	Randaug	Res&Crop	Stostich	Mixup	Cutout	Warm-up	LFW	CFP-FP	AgeDB-30	IJB-C		
1	√									99,63	95,72	97,1	95,29
2	√	95,87	√							99,68	96,84	97,55	
3	√	95,98	√	√						99,70	97,23	97,26	
4	√	96,05	√		√					99,73	97,40	97,30	
5	√	6	√		√	√				99,76	98,19	97,60	96,13
√	7	√	√		√		√			99,78	98,37	97,67	96,23
			√	√	√	√	√	√	√	99,80	98,78	97,85	96,37

Bảng 6: Tác động của việc tăng cường dữ liệu

B.2.2 Mức độ chồng chéo

Chúng tôi cũng kiểm tra mức độ chồng chéo của các bản vá được đào tạo bởi mạng của chúng tôi, chúng tôi tính toán tỷ lệ chồng chéo trung bình và phương sai của các bản vá gần nhất, được liệt kê trong Bảng 7. Tỷ lệ chồng chéo cho các tập dữ liệu tư thế lớn CFP-FP & IJB-C cao hơn so với các tập dữ liệu khác.

	LFW	CFP-FP	TuổiDB-30	IJB-C
R=16	0,5007±0,0002	0,5250±0,0016	0,4980±0,0002	0,5099±0,0007
R=49	0,3993±0,0002	0,4665±0,0064	0,3997±0,0001	0,4279±0,0003
R=196	0,2681±0,0001	0,2950±0,0010	0,2684±0,00008	0,2789±0,0005

Bảng 7: Tỷ lệ chồng chéo của các mảng lân cận thu được bởi phần fVIT-B của chúng tôi với R=16, 49 và 196

B.2.3 Ảnh hưởng của các CNN mốc khác nhau

Chúng tôi đã tiến hành một thí nghiệm để đánh giá tác động của việc sử dụng các CNN khác nhau cho các mốc quan trọng mạng. Cụ thể, chúng tôi cũng đã chọn Resnet-50[19]. Mô hình được sử dụng là phần fViT-B, với R = 196 điểm mốc. Bảng 8 cho thấy kết quả thu được. Chúng tôi kết luận rằng một điểm mốc lớn hơn CNN không tăng thêm độ chính xác cuối cùng.

Mạng lưới Landmark	LFW	CFP-FP	AgeDB	IJB-C	
fViT	99,85	98,13	99,21	98,29	97,29
phần fViT (MobilenetV3)	99,83	phần fViT	99,21	98,29	97,29
(ResNet50)	99,81		99,14	98,35	97,11

Bảng 8: Tác động của CNN mốc đến độ chính xác nhận dạng khuôn mặt.

B.3 Ứng dụng vào việc khám phá mốc không giám sát

Chúng tôi đã lựa chọn đánh giá định lượng các điểm mốc trên khuôn mặt được phát hiện bởi điểm mốc của chúng tôi CNN sử dụng giao thức đánh giá và cơ sở mã của [47]. Cụ thể, chúng tôi tuân theo [47] và báo cáo cái gọi là lỗi chuyển tiếp trên toàn bộ tập dữ liệu MAFL & AFLW trong Bảng 9.

lỗi chuyển tiếp là thước đo độ ổn định của mốc, đường ống của nó là đào tạo một hồi quy với các điểm mốc được dự đoán là dữ liệu đào tạo và 5 điểm mốc được dán nhãn thủ công trên MAFL & Bộ dữ liệu AFLW làm bộ thử nghiệm. Các điểm mốc dự đoán càng ổn định thì chúng càng tốt bản đồ đến sự thật cơ bản (để biết chi tiết và định nghĩa lỗi chuyển tiếp, vui lòng xem[47]). Vì nó có thể được quan sát phương pháp của chúng tôi cung cấp kết quả cạnh tranh với các phương pháp được đề xuất gần đây được thiết kế riêng cho mục đích xác định vị trí địa danh không cần giám sát.

	Phương pháp	MAFL AFLW	
Được giám sát	TCCN [76]	7,95	7,65
	MTCNN [75]	5.39	6,90
Không giám sát	Thewlis [54]	7.15	-
	Jakab [24]	3.19	6,86
	Trương [74]	3.46	7.01
	Thư [51]	5,45	-
	Sahasrabudhe [46]	6.07	-
	Sánchez [47]	3,99	6,69
	Mallis [40]	4.12	7.37
	Lý [34]	3.08	6.20
Của chúng tôi	Cột mốc CNN	4,87	10.22
	CNN mang tính bước ngoặt (R = 49)	3,37	7.16
	CNN mang tính bước ngoặt (R = 16)	3,88	7.69

Bảng 9: So sánh về khám phá mốc không giám sát. Kết quả lỗi chuyển tiếp [47] là đã báo cáo về toàn bộ tập dữ liệu MAFL & AFLW.