TransFace: Hiệu chỉnh đào tạo Transformer để nhận dạng khuôn mặt từ Góc nhìn tập trung vào dữ liệu

Quân Đan *1, Dương Lưu *2, Haoyu Xie2 , Đặng Kiến Khang3 ,
Hạo Nhiên Tạ4 , Xuansong Xie2 , Bạch Quý Tôn †2
1Đại học Chiết Giang 2Học viện Alibaba DAMO 3Đại học Hoàng gia London 4Đại học Thanh Hoa
danjun@zju.edu.cn,j.deng16@imperial.ac.uk,xiehr20@mails.tsinghua.edu.cn
{ly261666, xiehaoyu.xhy,xingtong.xxs, baigui.sbg}@alibaba-inc.com

Tóm tắt

Vision Transformers (ViTs) đã chứng minh khả năng biểu diễn mạnh mẽ trong nhiều nhiệm vụ trực quan khác nhau nhờ bản chất đói dữ liệu cố hữu của chúng. Tuy nhiên, chúng tôi bất ngờ phát hiện ra rằng ViT hoạt động yếu kém khi áp dụng cho khuôn mặt các tình huống nhận dạng (FR) với các tập dữ liệu cực lớn. Chúng tôi điều tra lý do của hiện tượng này và phát hiện ra rằng phương pháp tăng cường dữ liệu hiện có và chiến lược khai thác mẫu cứng không tương thích với xương sống FR dựa trên ViTs do thiếu sự cân nhắc phù hợp về việc bảo tồn thông tin cấu trúc khuôn mặt và tận dụng mỗi thông tin mã thông báo cục bộ. Để khắc phục những vấn đề này, bài báo này đề xuất một mô hình FR vượt trội được gọi là TransFace, sử dụng chiến lược tăng cường dữ liệu cấp bản vá được đặt tên là DPAP và một chiến lược khai thác mẫu cứng được đặt tên là EHSM. Đặc biệt, DPAP làm nhiễu loạn ngẫu nhiên biên độ thông tin về các bản vá lỗi chiếm ưu thế để mở rộng sự đa dạng của mẫu, giúp giảm thiểu hiệu quả vấn đề quá khớp trong ViTs. EHSM sử dụng entropy thông tin trong các mã thông báo cục bộ để điều chỉnh trọng số quan trọng một cách động của các mẫu dễ và khó trong quá trình đào tạo, dẫn đến một dự đoán ổn định hơn. Các thí nghiệm trên một số chuẩn mực chứng minh tính ưu việt của TransFace. Mã và các mô hình có sẵn tại https://github.com/DanJun6737/TransFace.

1. Giới thiệu

Trong vài năm qua, Mạng nơ-ron tích chập (CNN) [23, 32] đã đạt được thành công đáng kể trong cộng đồng thị giác máy tính, nhờ vào sự sẵn có của các tập dữ liệu quy mô lớn. Gần đây, sự ra đời của Vision Máy biến áp (ViTs) [12] đã thu hút sự chú ý của cộng đồng thị giác máy tính do khả năng biểu diễn mạnh mẽ của chúng. Không giống như các mô hình CNN, ViTs thiếu một số các độ lệch cảm ứng giống như tích chập, chẳng hạn như sự tương đương dịch chuyển và tính cục bộ, dẫn đến những thách thức trong quá trình hội tụ của xương sống ViTs. Để khắc phục vấn đề này, các tác phẩm pi-oneering [61, 12, 4, 82] chỉ ra lý do đằng sau điều này xuất phát từ bản chất đói dữ liệu của nó, cho thấy rằng một đại diện ViTs cao cấp sẽ được hỗ trợ bởi dữ liệu đào tạo quy mô lớn. Bằng cách tận dụng đặc tính cần dữ liệu nội tại của chúng, ViT thường được sử dụng để phục vụ như một xương sống thay thế cho một số nhiệm vụ trực quan [44, 12, 61, 58, 15, 17].

Tuy nhiên, khi xem xét một dữ liệu cực kỳ đầy đủ kịch bản để thỏa mãn tính chất đói dữ liệu của ViTs, cụ thể là Face Nhận dạng (FR), chúng tôi bất ngờ phát hiện ra rằng hiệu suất của ViT gần như bằng với hiệu suất của CNN [80]. Để khám phá lý do tại sao ViTs hoạt động yếu kém trong lĩnh vực FR, chúng tôi điều tra quá trình đào tạo của ViTs. Từ một dữ liệu tập trung Theo quan điểm này, chúng tôi phát hiện ra rằng phương pháp tăng cường dữ liệu cấp độ thể hiện và chiến lược khai thác mẫu cứng là không tương thích với xương sống FR dựa trên ViTs do thiếu của việc cân nhắc phù hợp về việc bảo tồn thông tin cấu trúc khuôn mặt và tận dụng từng thông tin mã thông báo cục bộ (minh họa trong Hình 1). Để xử lý những nhược điểm này, chúng tôi thực hiện hai nỗ lực sau đây: (i) Xác định lý do và cách thức thiết kế một chiến lược tăng cường dữ liệu cấp độ bản vá trên ViTs Xương sống FR. (ii) Tiết lộ lý do và cách khai thác thông tin mã thông báo đại diện.

Chiến lược tăng cường dữ liệu cấp bản vá. Do thiếu của các thành kiến quy nạp, các mô hình dựa trên ViT khó đào tạo và dễ bị quá khớp [61, 12, 4]. Để giảm bớt tình trạng quá khớp hiện tượng, các tác phẩm hiện có [61, 82, 70] cố gắng một số chiến lược tăng cường dữ liệu, chẳng hạn như Xóa ngẫu nhiên [81], Mixup [76], CutMix [75], RandAugment [7] và các biến thể của chúng [4, 19, 61, 38], để xây dựng các mẫu đào tạo đa dạng. Tuy nhiên, các chiến lược tăng cường dữ liệu cấp độ trường hợp này không phù hợp với nhiệm vụ FR, vì chúng chắc chắn sẽ

* Đóng góp ngang nhau, † Tác giả liên hệ ·

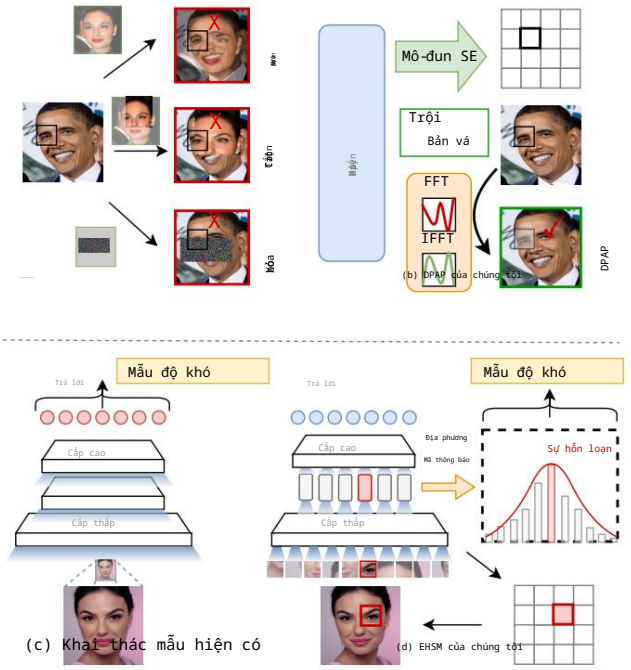
phá hủy một số thông tin cấu trúc quan trọng của nhận dạng khuôn mặt (như được hiển thị ở đầu Hình 1), điều này có thể dẫn đến ViTs tối ưu hóa theo hướng không chính xác. Hơn nữa, gần đây nghiên cứu [82] quan sát thấy rằng ViT thực sự có xu hướng quá phù hợp với một số bản vá cục bộ trong quá trình đào tạo, dẫn đến làm giảm nghiêm trọng hiệu suất tổng quát của mô hình. Ví dụ, trong nhiệm vụ FR, dự đoán của ViT có thể là chủ yếu là một vài mảng da trên mặt (ví dụ: mắt và trán). Do đó, một khi các mảng chính này bị xáo trộn (ví dụ, một siêu sao đeo kính râm hoặc đội mũ), mô hình có xu hướng đưa ra những quyết định sai lầm. Những vấn đề này ảnh hưởng nghiêm trọng việc triển khai trên diện rộng các mô hình FR dựa trên ViT trong thực tế các tình huống.

Để giải quyết vấn đề nói trên, được thúc đẩy bởi thông tin cấu trúc bảo toàn tính chất của Fourier phổ pha [50, 52, 73, 49], chúng tôi giới thiệu một cấp độ bản vá chiến lược tăng cường dữ liệu có tên là nhiễu loạn biên độ vá lỗi thống trị (DPAP). Không phá hủy độ trung thực và thông tin cấu trúc của khuôn mặt, DPAP có thể hiệu quả mở rộng sự đa dạng của mẫu. Cụ thể, DPAP sử dụng mô-đun Bóp và Kích thích (SE) [24] để sàng lọc ra top-K các bản vá (các bản vá chiếm ưu thế), sau đó trộn ngẫu nhiên các bản vá của chúng thông tin biên độ và kết hợp nó với thông tin gốc thông tin pha để tạo ra các mẫu đa dạng.

Khác với các chiến lược tăng cường dữ liệu trước đây, DPAP được đề xuất khéo léo sử dụng kiến thức trước đó (tức là, vị trí của các bản vá lỗi chiếm ưu thế) được cung cấp bởi mô hình để tăng cường dữ liệu, có thể làm giảm chính xác hơn vấn đề quá khớp trong ViTs. Hơn nữa, khi các bản vá đa dạng được tạo ra liên tục, DPAP cũng gián tiếp khuyến khích ViTs sử dụng các miếng dán mặt khác, đặc biệt là một số các bản vá dễ bị mạng sâu bỏ qua (ví dụ: tai, miệng và mũi) để đưa ra quyết định tự tin hơn.

Chiến lược khai thác mẫu cứng. Như đã trình bày trong Tài liệu tham khảo. [36, 26, 3], công nghệ khai thác mẫu cứng đóng vai trò quan trọng trong việc thúc đẩy hiệu suất cuối cùng của mô hình thông qua việc liên tục đồng hóa kiến thức từ các mẫu hiệu quả/cứng. Hầu hết các công trình trước đây được thiết kế đặc biệt cho CNN, họ thường áp dụng một số chỉ số cấp độ trường hợp của mẫu, chẳng hạn như xác suất dự đoán [36, 27], dự đoán mất mát [14, 56], và các tính năng tiềm ẩn [53], để khai thác các mẫu cứng (như thể hiện ở dưới cùng của Hình 1). Tuy nhiên, gần đây nghiên cứu [82] đã chỉ ra rằng dự đoán của ViT chủ yếu là chỉ được xác định bằng một vài mã thông báo vá lỗi, điều đó có nghĩa là mã thông báo toàn cầu của ViTs có thể bị chi phối bởi một số địa phương token. Do đó, trực tiếp sử dụng các chỉ số thiên vị như vậy để mẫu cứng của tôi không tối ưu cho ViTs, đặc biệt là khi một số mã thông báo địa phương chiếm ưu thế bị bỏ qua.

Để khai thác tốt hơn các mẫu cứng, lấy cảm hứng từ lý thuyết thông tin [51, 55, 5], chúng tôi đề xuất một mẫu cứng mới chiến lược khai thác có tên là Entropy-guided Hard Sample Min-ing (EHSM). EHSM coi ViT như một hệ thống xử lý thông tin, điều chỉnh động tầm quan trọng



Hình 1. (Trên cùng) Các phương pháp tăng cường dữ liệu trước đây có thể phá hủy độ trung thực và thông tin cấu trúc của nhận dạng khuôn mặt khi mẫu tăng cường. Chiến lược DPAP của chúng tôi không chỉ xây dựng các mẫu đa dạng mà còn bảo tồn hiệu quả thông tin chính của khuôn mặt. (Dưới cùng) Các phương pháp khai thác mẫu cứng hiện có thường áp dụng một số chỉ số cấp độ trường hợp để đo độ khó của mẫu, điều này không tối ưu đối với ViT. Chiến lược EHSM của chúng tôi tận dụng entropy thông tin từ tất cả các mã thông báo cục bộ để khai thác các mẫu cứng.

trọng lượng của các mẫu dễ và khó về mặt tổng số lượng của thông tin chứa trong các mã thông báo cục bộ. Nó có giá trị đề cập rằng EHSM có tiềm năng khuyến khích ViT khai thác hoàn toàn thông tin chi tiết có trong mỗi miếng vá trên khuôn mặt, đặc biệt là một số dấu hiệu khuôn mặt ít được chú ý (ví dụ, môi và hàm), giúp tăng cường đáng kể sức mạnh biểu diễn tính năng của mỗi mã thông báo cục bộ (như đã được xác minh bởi thí nghiệm trong Hình 2 và Hình 5). Theo cách này, ngay cả khi một số các bản vá hình ảnh quan trọng bị phá hủy, mô hình cũng có thể tận dụng tối đa các tín hiệu khuôn mặt còn lại để khái quát hóa mã thông báo toàn cầu, dẫn đến dự đoán ổn định hơn.

Sau đây là những đóng góp chính của bài báo này:

(1) Một chiến lược tăng cường dữ liệu cấp bản vá được đặt tên DPAP được giới thiệu để giảm thiểu hiệu quả tình trạng quá khớp vấn đề ở ViTs.

(2) Một chiến lược khai thác mẫu cứng mới có tên là EHSM được đề xuất để tăng cường tính ổn định của dự đoán mô hình FR.

(3) Kết quả thử nghiệm trên nhiều chuẩn mực khuôn mặt phổ biến đã cho thấy tính ưu việt của phương pháp của chúng tôi, ví dụ, chúng tôi đạt được độ chính xác 97,61% trên “TAR@FAR=1E-4” của

Đánh giá chuẩn IJB-C bằng bộ dữ liệu đào tạo Glint360K.

2. Các tác phẩm liên quan

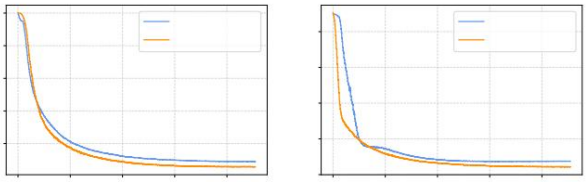
Vision Transformer (ViT). Gần đây, ViT đã chứng minh khả năng biểu diễn tính năng mạnh mẽ của mình trong nhiều tác vụ thị giác khác nhau, bao gồm nhận dạng hình ảnh[12, 61, 62, 44], phân đoạn ngữ nghĩa[58, 4] và phát hiện/định vị đối tượng [15, 17]. Không giống như CNN, ViT chủ yếu dựa trên về cơ chế tự chú ý [63], có thể hiệu quả nắm bắt mối quan hệ giữa các tính năng khác nhau. Nó có đã được chứng minh rằng ViT không tổng quát tốt khi được đào tạo trên số lượng mẫu đào tạo không đủ [12]. DeiT [61] giới thiệu một quy trình chưng cất kiến thức mới để tăng cường khả năng khái quát hóa của ViT. Để tốt hơn trích xuất cả thông tin hình ảnh toàn cầu và cục bộ, TNT [22] sử dụng một khối biến áp bên trong để mô hình hóa mối quan hệ giữa các bản vá phụ. ATS [16] được đề xuất để thiết kế các mô hình ViT hiệu quả về mặt tính toán bằng cách lấy mẫu các mã thông báo quan trọng một cách thích ứng.

Hơn nữa, một số nghiên cứu được đề xuất gần đây nhằm mục đích cân bằng tính toán và độ chính xác của ViT. UniFormer [34] hợp nhất tuyệt vời giữa phép tích chập 3D và sự tự chú ý không gian-thời gian, làm giảm đáng kể gánh nặng tính toán trong việc nắm bắt mối quan hệ mã thông báo. Dilateformer [29] sử dụng cửa sổ trượt để chọn các bản vá đại diện, giảm thiểu chi phí tính toán của sự chú ý bản thân một cách tuyệt vời. Tham khảo [30] đề xuất áp dụng cơ chế che giấu vào sự chú ý bản đồ, giảm đáng kể tải tính toán giữa mã thông báo.

Nhận dạng khuôn mặt (FR). CNN đã thực hiện đáng kể tiến triển trong các nhiệm vụ liên quan đến khuôn mặt [43, 40, 42, 11, 10, 41]. Trong số đó, việc trích xuất nhúng mặt sâu thu hút nhiều sự chú ý của các nhà nghiên cứu. Có hai cách chính để đào tạo CNN cho FR. Một loại cách là học tập dựa trên số liệu phương pháp, nhằm mục đích học cách biểu diễn khuôn mặt phân biệt, chẳng hạn như mất mát Triplet [53], mất mát Tuplet [57] và Trung tâm mất mát [68]. Một cách khác là các phương pháp softmax dựa trên biên độ, tập trung vào việc kết hợp hình phạt biên độ vào softmax khung phân loại mất mát, chẳng hạn như ArcFace [11, 9], Cos-Face [65], AM-softmax [64] và SphereFace [39]. Để cải thiện hơn nữa hiệu quả của mất mát softmax dựa trên biên độ trên tập dữ liệu quy mô lớn, trọng tâm của một số nghiên cứu đã thay đổi đến các tham số thích ứng [78, 77, 37, 31, 47], quy tắc hóa giữa các lớp [79, 13], khai thác mẫu [27, 67], tăng tốc học tập [1, 35], chắt lọc kiến thức [26], v.v.

Bộ chuyển đổi khuôn mặt được đề xuất gần đây [80] đầu tiên chứng minh tính khả thi của việc sử dụng ViT trong FR. Tuy nhiên, có vẫn còn thiếu sự khám phá về cách đào tạo mô hình FR dựa trên ViT vượt trội trên tập dữ liệu có quy mô cực lớn.

Data-Centirc ViTs. Để cải thiện hiệu suất của ViTs, các công trình trước đây [72, 22, 69, 74] chủ yếu cố gắng sửa đổi cấu trúc của các mô hình, điều này phụ thuộc rất nhiều về kinh nghiệm. Gần đây, một số tác phẩm đã được pro-



(Một) (b)

Hình 2. Chúng tôi hình dung xu hướng của entropy thông tin trung bình có trong mã thông báo cục bộ trong quá trình đào tạo. Với sự trợ giúp của EHSM, thông tin khuôn mặt chứa trong mỗi miếng vá là nhiều hơn được khai thác và sử dụng đầy đủ. Có thể tìm thấy thêm kết quả trong Hình 5.

được đưa ra để tăng cường khả năng tổng quát của ViTs từ quan điểm của dữ liệu đào tạo. Gong et al. [18] đề xuất hai các hàm mất mát liên quan đến bản vá để giảm bớt sự làm mịn quá mức vấn đề trong ViTs, có thể thúc đẩy quá trình đào tạo ổn định của ViTs sâu hơn mà không có bất kỳ sửa đổi cấu trúc nào. DeiT [61] cho thấy các chiến lược tăng cường dữ liệu mạnh mẽ [76, 81, 7, 6] có thể giúp ViTs hấp thụ dữ liệu hiệu quả hơn. TransMix [4] giới thiệu một chiến lược tăng cường dữ liệu tận dụng bản đồ chú ý của bộ chuyển đổi để hướng dẫn việc trộn nhãn. TokenMix [38] đề xuất trộn hình ảnh ở cấp độ mã thông báo, giúp ViTs suy ra các lớp mẫu chính xác hơn.

3. Phương pháp luận

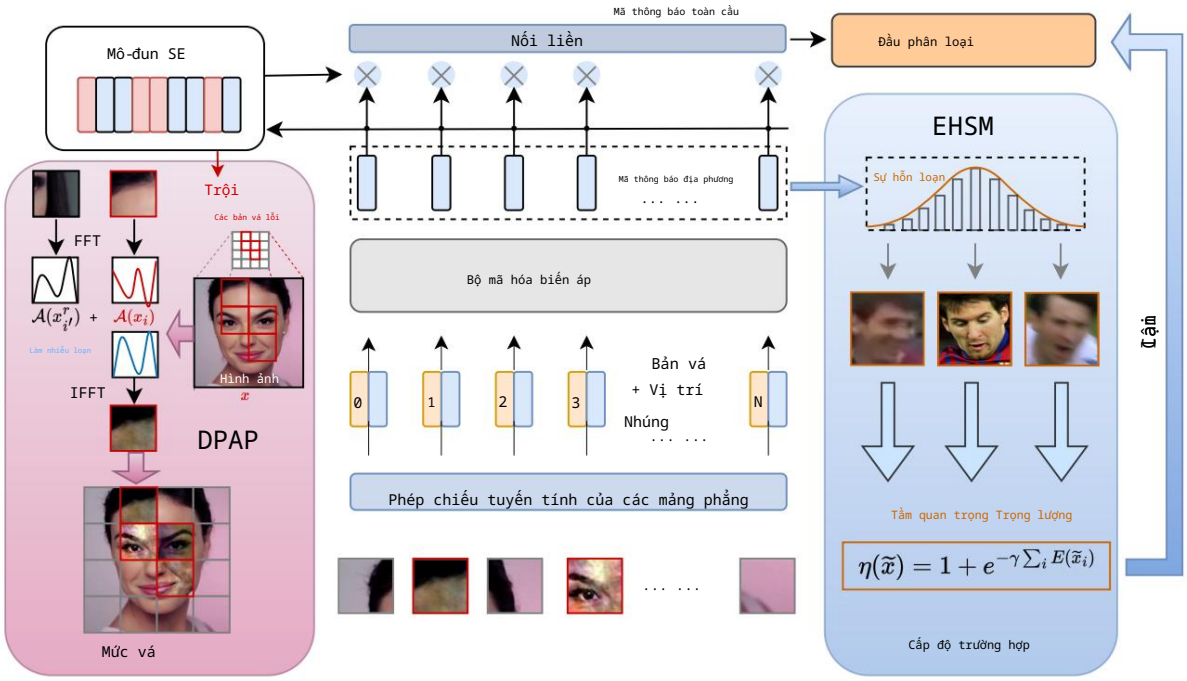
3.1. Chuẩn bị

Một ViT cổ điển [12] đầu tiên chia một hình ảnh đầu vào thành một chuỗi các bản vá có kích thước cố định. Mỗi bản vá nhỏ được ánh xạ tới một vectơ đặc trưng (hay còn gọi là mã thông báo) theo một lớp tuyến tính. Sau đó, một mã thông báo lớp có thể học được bổ sung được nối vào các mã thông báo và nhúng vị trí được thêm vào mỗi mã thông báo để giữ lại thông tin vị trí. Sau đó, các nhúng hỗn hợp được gửi vào bộ mã hóa biến áp để mã hóa tính năng. Cụ thể hơn, bộ mã hóa biến áp tiêu chuẩn bao gồm các khối tự chú ý nhiều đầu (MSA) và MLP.

LayerNorm (LN) và các kết nối còn lại được áp dụng trước và sau mỗi khối, tương ứng. Cuối cùng, lớp to-ken của đầu ra bộ mã hóa biến áp được chọn làm kết quả cuối cùng biểu diễn và đưa vào đầu phân loại để dự đoán.

Khác với kiến trúc ViT ban đầu, chúng tôi tuân theo insightface1 và không sử dụng mã thông báo lớp có thể học được trong mô hình của chúng tôi. Kiến trúc của mô hình TransFace của chúng tôi được thiết kế được mô tả trong Hình 3. Nó bao gồm ba thành phần: bộ mã hóa chuyển đổi F, mô-đun “Ép và Kích thích” (SE) S [24] và đầu phân loại C. Chúng tôi sẽ trình bày chi tiết kiến trúc mô hình của chúng tôi trong phần 3.2.

1https://github.com/deepinsight/insightface/tree/master/recognition



Hình 3. Tổng quan toàn cầu về mô hình TransFace được đề xuất của chúng tôi. Để giảm bớt vấn đề quá khớp trong ViT, chiến lược DPAP sử dụng Mô-đun SE để sàng lọc các mảng chiếm ưu thế top-K, sau đó làm nhiễu ngẫu nhiên thông tin biên độ của chúng để mở rộng tính đa dạng của mẫu. Hơn nữa, để khai thác tốt hơn các mẫu cứng và tăng cường sức mạnh trình bày tính năng của các mã thông báo cục bộ, chiến lược EHSM sử dụng cơ chế trọng số nhận biết entropy để cân nhắc lại tổn thất phân loại. n là tổng số bản vá và biểu thị phép nhân hoạt động giữa mã thông báo cục bộ và hệ số tỷ lệ được tạo ra bởi mô-đun SE. Các bản vá hình ảnh có hộp màu đỏ biểu thị sự thống trị miếng vá.

3.2. Chiến lược tăng cường dữ liệu cấp độ bản vá

Các tác phẩm hiện có chủ yếu áp dụng một loạt các cấp độ thể hiện các chiến lược tăng cường dữ liệu để giảm bớt hiện tượng quá khớp trong ViTs. Tuy nhiên, các chiến lược này chắc chắn sẽ phá hủy một số thông tin cấu trúc quan trọng của nhận dạng khuôn mặt (như được thể hiện trong Hình 1), có thể ảnh hưởng nghiêm trọng đến việc học của các biểu tượng phân biệt khuôn mặt. Hơn nữa, nghiên cứu gần đây [82] quan sát thấy rằng ViTs thực sự có xu hướng quá phù hợp với một một số bản vá lỗi, hạn chế đáng kể khả năng triển khai và mở rộng của các mô hình FR dựa trên ViT trong các tình huống ứng dụng.

Để giải quyết các vấn đề đã đề cập ở trên, một chiến lược tăng cường dữ liệu cấp bản vá mới có tên là Dominant Patch Amplitude Perturbation (DPAP) được đề xuất cho xương sống FR dựa trên ViTs. Các bước chính của chiến lược là: (1) Đầu tiên, chúng ta chèn một mô-đun SE S vào đầu ra của bộ mã hóa biến áp F và sử dụng các hệ số tỷ lệ được tạo ra bởi S để tìm ra các bản vá K hàng đầu (tức là, chiếm ưu thế các bản vá) của hình ảnh gốc x đóng góp nhiều nhất cho dự đoán cuối cùng. (2) Thứ hai, chúng tôi sử dụng một sự pha trộn tuyến tính cơ chế làm nhiễu loạn ngẫu nhiên thông tin biên độ của các bản vá lỗi chiếm ưu thế này. (3) Cuối cùng, chúng tôi đưa hình ảnh được tái tạo x vào mô hình TransFace của chúng tôi để giám sát đào tạo.

Về mặt toán học, cho một hình ảnh x, chúng ta biểu thị một chuỗi của các mảng hình ảnh phân hủy như x = (x1, x2, · · · , xn), trong đó mỗi xi biểu diễn một bản vá hình ảnh và n là tổng số số lượng bản vá. Và đầu ra của bộ mã hóa biến áp F được ký hiệu là (f1, f2, · · · , fn), trong đó f1, · · · , fn đại diện cho các mã thông báo cục bộ. Sau đó, tất cả các mã thông báo cục bộ được trích xuất bởi F sẽ đi qua mô-đun SE S và được định cỡ lại thành (κ1 · f1, κ2 · f2, · · · κn · fn), trong đó κ1, · · · biểu thị các hệ số định cỡ do S tạo ra. Trên thực tế, những , κn các hệ số tỷ lệ κ1, · · · κn của các mã thông báo cục bộ phản ánh gián tiếp tầm quan trọng của các mã thông báo cục bộ trong dự đoán. Chúng tôi tiếp tục chuẩn hóa các hệ số tỷ lệ này bằng cách sử dụng hàm softmax:

(1)

Theo các hệ số tỷ lệ chuẩn hóa lớn nhất của top-K, chúng ta có thể sàng lọc ra các bản vá chiếm ưu thế top-K mà mô hình "quan tâm đến" nhiều nhất. Để giải thoát ViTs khỏi quá phù hợp với các bản vá lỗi chiếm ưu thế này, một ý tưởng tự nhiên là cho phép mô hình "nhìn thấy" nhiều bản vá đa dạng hơn. Được thúc đẩy bởi tính chất bảo toàn thông tin cấu trúc của Fourier phổ pha [50, 52, 73, 49, 71], chúng tôi đề xuất để nhiễu loạn thông tin phổ biên độ của chúng

các bản vá sử dụng cơ chế trộn tuyến tính và giữ nguyên pha của chúng thông tin quang phổ không thay đổi.

Đối với một bản vá hình ảnh kênh đơn xi , Biến đổi Fourier T (xi) của nó có thể được biểu thị như sau:

(2)

nơi j 2 = 1, H và W biểu diễn chiều cao và chiều rộng của xi , tương ứng. T 1 (xi) biểu thị tương ứng trong câu thơ Biến đổi Fourier ánh xạ biên độ và pha thông tin trở lại không gian hình ảnh ban đầu. Fourier phép biến đổi và phép biến đổi ngược của nó có thể được thực hiện hiệu quả bằng các thuật toán FFT và IFFT [2].

A(xi) và P(xi) biểu diễn phổ biên độ và phổ pha của miếng vá hình ảnh xi , tương ứng:

(3)

trong đó R(xi) và I(xi) biểu thị phần thực và phần ảo của T (xi), tương ứng. Đối với các bản vá hình ảnh khuôn mặt RGB, chúng tôi cần tính toán Biến đổi Fourier của mỗi kênh độc lập để có được phổ biên độ cuối cùng và phổ pha.

Để xây dựng hiệu quả các hình ảnh đa dạng mà không phá hủy độ trung thực và thông tin cấu trúc của nhận dạng khuôn mặt, chúng tôi sử dụng cơ chế giống như Mixup để trộn tuyến tính phổ biên độ của bản vá ưu thế xi và bản vá ngẫu nhiên x r :

(4)

trong đó λ U(0, α), U(0, α) là phân phối đồng đều trên [0, α], x r là các bản vá ngẫu nhiên của một mẫu đào tạo ngẫu nhiên x r , và α là siêu tham số được sử dụng để kiểm soát cường độ của thông tin biên độ pha trộn. Sau đó, chúng tôi kết hợp phổ biên độ hỗn hợp với phổ pha ban đầu để tái tạo một biểu diễn Fourier mới:

(5)

sẽ được ánh xạ vào không gian hình ảnh gốc bởi biến đổi Fourier ngược để tạo ra một bản vá mới, tức là xi = T 1 [T (xi)(u, v)]. Sau đó chúng tôi đưa hình ảnh tăng cường x vào mô hình đào tạo có giám sát.

Đối với hình ảnh tăng cường x, chúng tôi biểu thị các to-ken cục bộ được trích xuất bởi F là (f1, f2, · · · , fn), sẽ là tiếp tục được đưa vào mô-đun SE S và được định cỡ lại thành (κ1 · κ2 · f2, · · · cal ,κn · fn). Sau đó, tất cả các mã thông báo lo- f1, được chia tỷ lệ lại được nối vào một mã thông báo toàn cục g = κ1 · f1; κ2 · f2; · · · ; κn · fn , sẽ được sử dụng cho nhiệm vụ FR tiếp theo thông qua đầu phân loại C. Trong

Hình 4. Hình ảnh ví dụ và thông tin entropy tương ứng. Các mẫu được dán nhãn có cùng ID sẽ được hiển thị ở mỗi cột. (Hàng đầu tiên) Các mẫu dễ thường chứa thông tin phong phú hơn (tức là, entropy thông tin lớn hơn). (Hàng thứ hai) Các mẫu cứng thường chứa ít thông tin hơn (tức là, entropy thông tin thấp hơn).

mô hình, chúng tôi áp dụng ArcFace Loss được sử dụng rộng rãi nhất [11] như là tổn thất phân loại cơ bản:

(6)

trong đó y biểu thị nhãn lớp của hình ảnh x, s là một tỷ lệ siêu tham số, θl là góc giữa cột thứ l trọng lượng và đặc điểm, m > 0 biểu thị biên độ góc cộng và c là số lớp.

Khi các mảng chiếm ưu thế liên tục thay đổi, Chiến lược DPAP có thể gián tiếp thúc đẩy mô hình FR để sử dụng các miếng vá mặt khác, đặc biệt là một số miếng vá dễ dàng bị bỏ qua bởi mạng lưới sâu (ví dụ, tai, miệng và mũi), để hỗ trợ dự đoán cuối cùng. Quan trọng hơn, chiến lược DPAP khéo léo sử dụng kiến thức trước đó được cung cấp bởi mô hình (tức là vị trí của người thống trị các bản vá) để thực hiện tăng cường dữ liệu, có thể giảm thiểu hiệu quả hơn vấn đề quá khớp và tăng cường khả năng khái quát của ViTs.

3.3. Chiến lược khai thác mẫu cứng dựa trên Entropy

Như được thể hiện trong Tài liệu tham khảo [36, 26, 3], công nghệ khai thác mẫu cứng thường được áp dụng để thúc đẩy hơn nữa quá trình cuối cùng của mô hình hiệu suất. Các công trình trước đây về khai thác mẫu cứng, chẳng hạn như như Mất tiêu cự [36], MV-Softmax [67], OHEM [56], ATk mất mát [14], v.v., được thiết kế đặc biệt cho CNN, chúng nhắm mục tiêu để khuyến khích các mô hình nhấn mạnh rõ ràng tác động của mẫu cứng [27, 53]. Các phương pháp này thường sử dụng một số chỉ số cấp độ trường hợp của mẫu, chẳng hạn như xác suất dự đoán [36, 27], mất mát dự đoán [14, 56] và tiềm ẩn các tính năng [53], để đo trực tiếp hoặc gián tiếp độ khó của mẫu. Tuy nhiên, nghiên cứu gần đây [82] đã chứng minh rằng dự đoán của ViT chủ yếu được xác định bởi chỉ một một số mã thông báo bản vá, có nghĩa là mã thông báo toàn cầu của ViT có thể bị chi phối bởi một số mã thông báo cục bộ. Do đó, đối với ViTs, việc sử dụng các chỉ số thiên vị như vậy để khai thác là không tối ưu mẫu cứng, đặc biệt là khi một số mã thông báo cục bộ chiếm ưu thế bị bỏ qua.

Để khai thác các mẫu cứng chính xác hơn, được thúc đẩy bởi lý thuyết thông tin [51, 55, 5], chúng tôi đề xuất đo lường độ khó của mẫu theo tổng lượng thông tin được chứa trong các mã thông báo cục bộ. Như được minh họa trong Hình 4, hình ảnh khuôn mặt chất lượng cao (mẫu dễ dàng) thường chứa thông tin phong phú hơn (tức là entropy thông tin cao hơn) và do đó dễ học hơn bằng mô hình. Khuôn mặt chất lượng thấp hình ảnh (mẫu cứng), chẳng hạn như hình ảnh khuôn mặt mờ và hình ảnh khuôn mặt có độ tương phản thấp, thường chứa ít thông tin hữu ích hơn thông tin (tức là, entropy thông tin thấp hơn), vì vậy chúng là khó học hơn.

Khi chúng ta coi mạng nơ-ron sâu M là một hệ thống xử lý thông tin, chúng ta có thể tóm tắt cấu trúc của nó thành đồ thị G = (Z, Q). Một loạt các nơ-ron tạo thành đỉnh đặt Z và các kết nối giữa các nơ-ron tạo thành cạnh đặt Q. Đối với bất kỳ z Z và q Q, e(z) và e(q) biểu thị giá trị của mỗi đỉnh z và mỗi cạnh q tương ứng. Do đó, không gian trạng thái liên tục của mạng sâu M có thể là được xác định bởi tập hợp Ω = {e(z), e(q) : z Z, q Q}. Trong theo cách này, tổng thông tin chứa trong M có thể được đo bằng entropy E(Ω) của tập Ω. Các tập E(Ωz) = {e(z) : z Z} và E(Ωq) = {e(q) : q Q} biểu diễn tổng thông tin chứa trong các tính năng tiềm ẩn và trong các tham số mạng, tương ứng. Đặc biệt, E(Ωz) đo lường sức mạnh biểu diễn đặc điểm của mạng M và E(Ωq) đo lường độ phức tạp của mạng. Trong công việc của chúng tôi, chúng tôi tập trung vào entropy của các tính năng tiềm ẩn E(Ωz) hơn là entropy của các tham số mạng E(Ωq).

Tuy nhiên, trong mạng sâu M, các tính năng tiềm ẩn của hình ảnh luôn tuân theo một phân phối phức tạp và chưa biết [20, 8], rất khó để tính toán trực tiếp thông tin entropy của các đặc điểm tiềm ẩn E(Ωz). May mắn thay, Nguyên lý Entropy cực đại [5, 28, 33, 60] đã chứng minh rằng entropy của một phân phối bị giới hạn trên bởi một Gaussian phân phối có cùng giá trị trung bình và phương sai, như thể hiện trong Định lý 1.

Định lý 1 Đối với bất kỳ phân phối liên tục D(a) của trung bình μ và phương sai σ ^{2}, entropy vi phân của nó được tối đa hóa khi D(a) là phân phối chuẩn Gauss N (μ, σ2).

Do đó, chúng ta có thể ước tính giới hạn trên của entropy thay thế. Giả sử a được lấy mẫu từ phân phối Gaussian N (μ, σ2), entropy vi phân của a có thể được xác định như sau:

— —

(7)

Như có thể thấy, entropy của phân phối Gaussian chỉ phụ thuộc vào phương sai. Theo cách này, chúng ta có thể xấp xỉ entropy của các đặc điểm tiềm ẩn E(Ωz) một cách hiệu quả chỉ bằng cách tính toán phương sai của các tính năng tiềm ẩn.

Khác với các công trình trước đây sử dụng các chỉ số thiên vị để khai thác mẫu cứng, chúng tôi đề xuất một phương pháp khai thác mẫu cứng mới.

Dữ liệu đào tạo	Phương pháp	GFLOP	LFW	CFP-FP	AgeDB-30
Glint360K	R50, Mặt hồ quang	6,3	99,78	98,77	98,28
	R100, Mặt hồ quang	12,1	99,81	99,04	98,31
	R200, Mặt hồ quang	23,4	99,82	99,14	98,49
	ViT-S	5,7	99,80	98,85	98,24
	ViT-B	11,4	99,82	99,02	98,33
	ViT-L	25,3	99,82	99,10	98,47
	TransFace-S	5,8	99,85	98,91	98,50
	TransFace-B	11,5	99,85	99,17	98,53
	TransFace-L	25,4	99,85	99,32	98,62

Bảng 1. Độ chính xác xác minh (%) trên các chuẩn mực LFW, CFP-FP và AgeDB-30.

chiến lược khai thác đơn giản có tên là Mẫu cứng được hướng dẫn theo Entropy Khai thác (EHSM) để đạt được mục tiêu này tốt hơn. EHSM xem xét toàn diện thông tin cục bộ và toàn cầu của các mã thông báo trong việc đo lường độ khó của mẫu. Cụ thể, đối với một mẫu tăng cường x = (x1, x2, · · · , xn), EHSM đầu tiên ước tính entropy thông tin cục bộ E(xi) = E(κi · fi) của mỗi mã thông báo cục bộ κi ·fi sử dụng Phương trình (7). Sau đó, tất cả các mã thông báo cục bộ entropy thông tin được tổng hợp như thông tin toàn cầu entropy E(xi) của mẫu x. Cuối cùng, EHSM sử dụng một cơ chế trọng lượng nhận thức entropy η(x) = 1 + e $^{E(xi)}$ $^{y tôi}$ để gán trọng số quan trọng một cách thích ứng cho mỗi mẫu, trong đó γ là hệ số nhiệt độ. Tổn thất phân loại được cân nhắc lại có thể được lập công thức như sau:

(8)

Điều đáng nói là EHSM khuyến khích rõ ràng mô hình tập trung vào các mẫu cứng có ít thông tin. Để giảm thiểu Lcls mục tiêu, mô hình phải tối ưu hóa cả trọng số η và mất phân loại cơ bản Larc trong quá trình đào tạo, điều này sẽ mang lại hai lợi ích: (1) Giảm thiểu Larc có thể khuyến khích mô hình học các đặc điểm khuôn mặt tốt hơn từ các mẫu đào tạo đa dạng. (2) Giảm thiểu trọng lượng η(x) (tức là, tối đa hóa tổng thông tin E(xi)) sẽ tạo điều kiện cho mô hình khai thác đầy đủ thông tin đặc điểm có trong mỗi mảng khuôn mặt, đặc biệt là một số khuôn mặt ít được chú ý các tín hiệu (ví dụ, mũi, môi và hàm), giúp tăng cường đáng kể sức mạnh biểu diễn tính năng của mỗi mã thông báo cục bộ. Trong này theo cách này, ngay cả khi một số miếng vá quan trọng trên khuôn mặt bị phá hủy, mô hình cũng có thể tận dụng tối đa các tín hiệu khuôn mặt còn lại để tổng quát hóa mã thông báo toàn cầu, dẫn đến dự đoán ổn định hơn.

4. Thí nghiệm

4.1. Chi tiết triển khai

Bộ dữ liệu. Chúng tôi áp dụng riêng bộ dữ liệu MS1MV2 [11] (5,8 triệu hình ảnh, 85 nghìn danh tính) và đề xuất gần đây bộ dữ liệu Glint360K [1] quy mô lớn hơn (17 triệu hình ảnh, 360K danh tính) để đào tạo mô hình của chúng tôi. Để đánh giá, chúng tôi sử dụng LFW[25], AgeDB-30 [48], CFP-FP [54] và IJB-C [46] như các tiêu chuẩn để kiểm tra hiệu suất nhận dạng của chúng tôi người mẫu.

Dữ liệu đào tạo	Phương pháp	GFLOP	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)	IJB-C(1E-3)	IJB-C(1E-2)	IJB-C(1E-1)
MS1MV2	R100, Softmax [47]	12,1*	64,07	83,68	92,40	-	-	-
	R100, SV-AM-Softmax [66, 47]	12,1	63,65	80,30	88,34	-	-	-
	R100, Mặt cầu [39, 47]	12,1	68,86	83,33	91,77	-	-	-
	R100, CosFace [65, 47]	12,1	87,96	92,68	95,56	-	-	-
	R100, Mặt hổ quang [11]	12,1	-	-	95,60	-	-	-
	R100, MV-Arc-Softmax [67, 27]	12,1*	-	-	95,20	-	-	-
	R100, Vòng tròn mất mát [59]	12,1	-	89,60	93,95	96,29	-	-
	R100, Khuôn mặt học tập [27]	12,1	-	-	96,10	-	-	-
	R100, Mặt Mag [47]	12,1	89,26	93,67	95,81	-	-	-
	ViT-S	5,7	86,14	93,40	95,89	97,24	98,21	98,80
	ViT-B	11,4	86,66	94,08	96,15	97,38	98,24	98,89
	ViT-L	25,3	86,77	94,11	96,24	97,42	98,26	98,94
	TransFace-S	5,8	86,75	93,87	96,45	97,51	98,34	98,99
	TransFace-B	11,5	86,73	94,15	96,55	97,73	98,47	99,11
	TransFace-L	25,4	86,90	94,55	96,59	97,80	98,45	99,04
Glint360K	R50, Mặt hổ quang	6,3	88,40	95,29	96,81	97,79	98,30	99,04
	R100, Mặt hổ quang	12,1	88,38	95,38	96,89	97,86	98,33	99,07
	R200, Mặt hổ quang	23,4	89,45	95,71	97,20	97,98	98,38	99,09
	ViT-S	5,7	88,52	95,24	96,70	97,71	98,29	99,01
	ViT-B	11,4	88,58	95,41	96,88	97,80	98,35	99,09
	ViT-L	25,3	89,69	95,78	97,13	97,91	98,43	99,09
	TransFace-S	5,8	89,93	96,06	97,33	98,00	98,49	99,11
	TransFace-B	11,5	88,64	96,18	97,45	98,17	98,66	99,23
	TransFace-L	25,4	89,71	96,29	97,61	98,26	98,64	99,19

Bảng 2. Độ chính xác xác minh (%) trên chuẩn IJB-C. biểu thị R100 GFLOP ở độ phân giải 112 × 112.

Cài đặt đào tạo. Các thí nghiệm của chúng tôi được thực hiện bằng cách sử dụng Pytorch trên 8 GPU NVIDIA Tesla V100. Chúng tôi theo dõi [11] sử dụng ArcFace (s = 64 và m = 0,5) làm phân loại cơ bản và cắt tất cả các hình ảnh đầu vào thành 112×112 bằng RetinaFace [10, 21]. Để tối ưu hóa các mô hình FR dựa trên ViT, chúng tôi áp dụng trình tối ưu hóa AdamW [45] với sự suy giảm trọng số của 0,1 để hội tụ tốt hơn. Tỷ lệ học tập cơ bản cho MS1MV2 được đặt thành 1e-3 và 1e-4 cho Glint360K. Kiến trúc chi tiết của các mô hình ViT có thể được tìm thấy trong insight-face2

· Mô-đun SE bao gồm hai lớp được kết nối đầy đủ, mỗi lớp có 144 nơ-ron, tiếp theo là ReLu và Sigmoid các hàm kích hoạt, tương ứng. Lưu ý rằng tất cả các mô hình được học từ đầu mà không cần đào tạo trước. Trong thực tế, chúng tôi áp dụng phương sai thay vì entropy để đo lượng của thông tin chứa trong mỗi mã thông báo cục bộ để tối ưu hóa mô hình ổn định hơn. Đối với siêu tham số α trong DPAP, chúng tôi chọn α = 1 cho tất cả các thí nghiệm.

4.2. Kết quả trên các chuẩn mực chính thống

Kết quả trên LFW, CFP-FP và AgeDB-30. Chúng tôi đào tạo TransFace trên Glint360K và so sánh nó với các phương pháp khác trên nhiều điểm chuẩn khác nhau, như được báo cáo trong Bảng 1. Chúng ta có thể thấy rằng hiệu suất của đường cơ sở ViT có thể so sánh được với hiệu suất của mô hình dựa trên ResNet. Đáng chú ý là độ chính xác của mô hình TransFace của chúng tôi trên các điểm chuẩn này đã gần bão hòa. Đặc biệt, TransFace-L cao hơn ViT-L tăng lần lượt là +0,03%, +0,22% và +0,15% trên ba tập dữ liệu.

Kết quả trên IJB-C. Chúng tôi đào tạo TransFace của mình trên MS1MV2 và Glint360K tương ứng, và so sánh với các đối thủ cạnh tranh của SOTA trên chuẩn IJB-C, như được báo cáo trong Bảng 2. Chúng ta có thể quan sát thấy ba mô hình TransFace của chúng tôi đã được đào tạo với tập dữ liệu MS1MV2 đã đánh bại đáng kể các tập dữ liệu dựa trên ResNet khác mô hình trên “TAR@FAR=1E-4”. Ví dụ, so sánh đối với CurricularFace, TransFace-B đạt được mức cải thiện +0,45% trên “TAR@FAR=1E-4”. Hơn nữa, TransFace-S vượt trội hơn ViT-S +0,56% trên “TAR@FAR=1E-4”.

Trên bộ đào tạo Glint360K, các mô hình của chúng tôi đáng kể vượt trội hơn các đối thủ cạnh tranh khác. Đặc biệt, TransFace-L đạt được kết quả tốt nhất chung và vượt trội hơn ViT-L rất nhiều với +0,48% và +0,51% trên “TAR@FAR=1E-4” và “TAR@FAR=1E-5”, tương ứng. Những kết quả cải thiện này chứng minh tính ưu việt của TransFace của chúng tôi.

Phương pháp	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)
ViT-S	86,14	93,40	95,89
ViT-S + SE	86,26	93,76	96,12
ViT-S + DPAP	86,60	93,82	96,30
TransFace-S	86,75	93,87	96,45

Bảng 3. Nghiên cứu cắt bỏ mô hình của chúng tôi. Dữ liệu đào tạo: MS1MV2.

4.3. Phân tích và Nghiên cứu cắt bỏ

1) Đóng góp của từng thành phần: Để điều tra đóng góp của từng thành phần trong mô hình của chúng tôi, chúng tôi sử dụng MS1MV2 là tập huấn luyện và so sánh TransFace-S, ViT-S (cơ sở) và hai biến thể của TransFace-S trên Tiêu chuẩn IJB-C. Các biến thể của TransFace-S như sau: (1) ViT-S + SE, biến thể chỉ giới thiệu SE mô-đun trong mô hình ViT-S. (2) ViT-S + DPAP, dựa trên ViT-S, biến thể bổ sung chiến lược DPAP.

2https://github.com/deepinsight/insightface/tree/master/recognition

Phương pháp	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)
ViT-S 86,14 ViT-S + Xóa ngẫu		93,40	95,89
nhiên 83,68 ViT-S + RandAugment 86,24 ViT-S		93,27	96,06
+ PatchErasing 86,26 85,51 ViT-S + Mixup		93,74	96,14
ViT-S + CutMix 85,30 ViT-S + DPAP 86,60		93,65	96,11
4. So sánh với các	Bảng	93,41	96,08
chiến lược tăng cường dữ liệu trước đây.		93,53	96,12
		93,82	96,30

Dữ liệu đào tạo: MS1MV2.

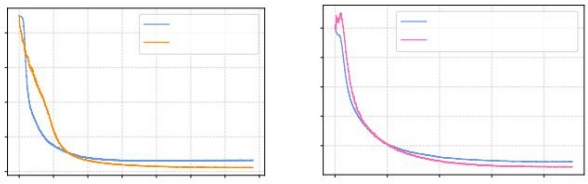
Phương pháp	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)
ViT-S + Tấn công	86,24	93,74	96.03
ViT-S + MV-Softmax	86,26	93,73	96,08
ViT-S + Mất tiêu cự	86,14	93,71	96,11
ViT-S + EHSM (toàn cầu)	86,41	93,76	96,13
ViT-S + EHSM	86,46	93,85	96,22

Bảng 5. So sánh với các chiến lược khai thác mẫu cứng trước đây.
Dữ liệu đào tạo: MS1MV2.

Phương pháp tham số K		IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)
	TransFace-S	85,28 93,72	86,30 93,72	96,38
1	TransFace-S	85,50 93,78	86,75 93,87	96,42
3	TransFace-S	86,30 93,72	84,90 93,65	96,43
5	TransFace-S	Bảng 6. Phân tích độ nhạy		96,45
7	TransFace-S	tham số của mô hình của		96,42
10 20	TransFace-S	chúng tôi trên IJB-C.		96,35

Dữ liệu đào tạo: MS1MV2.

Kết quả thu thập được trong Bảng 3 phản ánh những quan sát sau: (1) So với ViT-S, độ chính xác của ViT-S
+ SE được cải thiện đôi chút do bổ sung mô-đun SE. (2) ViT-S + DPAP hoạt động tốt hơn ViT-S + SE, điều này chỉ ra rằng việc làm nhiễu phổ biên độ của ưu thế
các bản vá có thể làm giảm hiệu quả vấn đề quá khớp trong ViTs. (3) TransFace-S hoạt động tốt hơn ViT-S + DPAP, điều này cho thấy hiệu quả của chiến lược EHSM của chúng tôi.
2) So sánh với dữ liệu tăng cường trước đó
Chiến lược: Để chứng minh thêm sự vượt trội của chúng tôi Chiến lược DPAP, chúng tôi so sánh nó với các chiến lược tăng cường dữ liệu hiện có, bao gồm Xóa ngẫu nhiên [81], Trộn lẫn [76], CutMix [75], RandAugment [7] và đề xuất gần đây PatchErasing [82]. Chúng tôi sử dụng MS1MV2 để đào tạo các mô hình và đánh giá hiệu suất của họ trên IJB-C. Như đã báo cáo trong Bảng 4, so với các chiến lược khác, chiến lược DPAP của chúng tôi có thể mang lại hiệu suất cao hơn cho ViT, được hưởng lợi từ việc sử dụng kiến thức trước đó (tức là vị trí
(các mảng chiếm ưu thế) và việc bảo tồn cấu trúc hình thành nên đặc điểm khuôn mặt.
3) EHSM có tăng cường khả năng biểu diễn tính năng không Sức mạnh của mỗi mã thông báo? Chúng tôi điều tra xu hướng của entropy thông tin trung bình có trong mã thông báo cục bộ của ViT (cơ sở) và biến thể ViT + EHSM trong quá trình đào tạo trên tập dữ liệu MS1MV2, như thể hiện trong Hình 2 và 5a. Chúng ta có thể thấy rằng với việc bổ sung chiến lược EHSM của chúng tôi, thông tin cấp mã thông báo trở nên phong phú hơn, điều này chứng minh tính ưu việt của chiến lược EHSM trong việc cải thiện tính năng



(Một) (b)

Hình 5. Xu hướng của entropy thông tin trung bình chứa trong mã thông báo cục bộ trong quá trình đào tạo. Với sự trợ giúp của EHSM, thông tin khuôn mặt chứa trong mỗi bản vá được khai thác đầy đủ hơn và được sử dụng.

Hình 6. (Hàng đầu tiên) Các mẫu đào tạo ban đầu. (Hàng thứ hai) Các mẫu đào tạo được tăng cường bằng chiến lược DPAP.

sức mạnh biểu diễn của mỗi mã thông báo cục bộ.
4) Hiệu quả của EHSM: Để chứng minh tính ưu việt của EHSM trong khai thác mẫu cứng, chúng tôi so sánh nó với các chiến lược trước đây, bao gồm mất ATk [14], MV-Softmax [67] và Mất tiêu cự [36], như được báo cáo trong Bảng 5. Chúng ta có thể quan sát thấy rằng chiến lược EHSM được đề xuất vượt trội hơn đáng kể so với các chiến lược khai thác mẫu cứng trước đây, điều này chỉ ra rằng chiến lược EHSM của chúng tôi có thể đo mẫu tốt hơn độ khó và tăng cường hiệu suất của mô hình.

Hơn nữa, để xác nhận những lợi thế của Chiến lược EHSM trong khai thác mẫu cứng bằng cách toàn diện tận dụng cả thông tin địa phương và toàn cầu, chúng tôi tiến hành so sánh thêm giữa ViT-S + EHSM và biến thể của nó ViT-S + EHSM (toàn cầu) sử dụng trực tiếp entropy của mã thông báo toàn cầu để đo độ khó của mẫu. Các kết quả thu thập được trong Bảng 5 chứng minh rằng ViT-S + EHSM vượt trội hơn nhiều so với biến thể ViT-S + EHSM (toàn cầu), chỉ ra rằng việc kết hợp đầy đủ entropy thông tin của tất cả các mã thông báo cục bộ có thể đo lường mẫu toàn diện hơn khó khăn hơn so với việc chỉ sử dụng entropy của mã thông báo toàn cầu. Hơn nữa, so với biến thể ViT-S + EHSM (toàn cầu), ViT-S + EHSM có thể tăng cường tính năng hiệu quả hơn sức mạnh biểu diễn của mỗi mã thông báo cục bộ, như minh họa trong Hình 5b.
5) Hình ảnh hóa DPAP: Như thể hiện trong Hình 6, chúng tôi hình ảnh hóa các mẫu đào tạo ban đầu và các mẫu được tăng cường bởi chiến lược DPAP (K = 15) trong quá trình đào tạo trên MS1MV2. Chúng ta có thể thấy rằng các bản vá lỗi chiếm ưu thế chủ yếu được phân phối gần tóc, trán và mắt, phù hợp với thị lực của chúng ta

học phí. Chiến lược DAPA được đề xuất có hiệu quả làm giảm
mô hình từ việc lắp quá mức vào các bản vá lỗi chiếm ưu thế này bằng
làm nhiễu loạn thông tin biên độ Fourier của chúng, gián tiếp khuyến khích
ViT sử dụng các tín hiệu khuôn mặt còn lại
(ví dụ, mũi, miệng, tai và hàm) để hỗ trợ dự đoán cuối cùng, tăng cường
đáng kể khả năng khái quát hóa của mô hình
khả năng.
6) Độ nhạy tham số: Để nghiên cứu tác động của tham số K (tức là số lượng
các mảng ưu thế được chọn)
trong mô hình của chúng tôi, chúng tôi áp dụng tập dữ liệu MS1MV2 để đào tạo
TransFace-S với K khác nhau và đánh giá hiệu suất của chúng trên IJB-C. Các
kết quả thu thập được trong Bảng 6 chỉ ra
rằng mô hình của chúng tôi mạnh mẽ với K. Khi K tăng, độ chính xác tổng
thể đầu tiên tăng lên rồi giảm xuống, điều này xác minh rằng
làm nhiễu loạn thông tin biên độ của sự thống trị
Các bản vá có thể làm giảm hiệu quả vấn đề quá khớp.

5. Kết luận

Trong bài báo này, chúng tôi phát triển một mô hình mới có tên là Trans-
Face để cứu vãn hiệu suất dễ bị tổn thương của ViT trong FR
nhiệm vụ. Đặc biệt, chúng tôi giới thiệu một chiến lược tăng cường dữ liệu
cấp bản vá có tên là DPAP và một chiến lược khai thác mẫu cứng có tên là
EHSM. Trong số đó, DPAP áp dụng một hỗn hợp tuyến tính
cơ chế làm nhiễu thông tin biên độ của các bản vá chiếm ưu thế để giảm
thiểu vấn đề quá khớp trong ViTs.
EHSM tận dụng hoàn toàn entropy thông tin của nhiều mã thông báo cục bộ để
đo độ khó của mẫu, cải thiện đáng kể
sức mạnh biểu diễn tính năng của các mã thông báo cục bộ. Ngoài
việc bổ sung mô-đun SE, TransFace không đưa ra bất kỳ thay đổi đáng kể nào
về mặt kiến trúc. Toàn diện
các thí nghiệm trên các chuẩn mực khuôn mặt phổ biến xác minh tính ưu việt của
TransFace. Chúng tôi hy vọng những phát hiện của chúng tôi có thể làm sáng tỏ một số
ánh sáng cho nghiên cứu trong tương lai về FR dựa trên ViTs cũng như một số
chủ đề có liên quan, ví dụ, tạo văn bản thành hình ảnh được cá nhân hóa
mô hình (AIGC) và tái tạo khuôn mặt 3D.

Tài liệu tham khảo

[1] Tương An, Đặng Kiến Khang, Giả Quốc, Tử Dung Phong, Hứa Hán
Zhu, Jing Yang và Tongliang Liu. Giết hai con chim bằng
one stone: Đào tạo nhận dạng khuôn mặt hiệu quả và mạnh mẽ
cnns của fc một phần. Trong Biên bản báo cáo của Hội nghị IEEE/
CVF về Tầm nhìn máy tính và Nhận dạng mẫu, các trang
4042–4051, 2022.

[2] E Oran Brigham và RE Morrow. Biến đổi Fourier nhanh.
Phổ IEEE, 4(12):63–70, 1967.

[3] Beidi Chen, Weiyang Liu, Zhiding Yu, Jan Kautz, Anshu-mali
Shrivastava, Animesh Garg và Animashree Anandku-mar. Độ cứng thị
giác góc cạnh. Trong Hội nghị quốc tế
về Học máy, trang 1637–1648. PMLR, 2020.

[4] Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan
Yuille và Song Bai. Transmix: Tham gia vào việc mix để có tầm nhìn
máy biến áp. Trong Biên bản báo cáo của Hội nghị IEEE/CVF
về Tầm nhìn máy tính và Nhận dạng mẫu, trang 12135–
12144, 2022.

[5] Thomas M Cover. Các yếu tố của lý thuyết thông tin. John
Wiley & Sons, 1999.

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasude-van và
Quoc V Le. Tự tăng cường: Tăng cường học tập
chính sách từ dữ liệu. bản in trước arXiv arXiv:1805.09501, 2018.

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens và Quốc 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 mẫu
hội thảo công nhận, trang 702–703, 2020.

[8] Jun Dan, Tao Jin, Hao Chi, Yixuan Shen, Jiawang Yu, và
Jinhai Zhou. Homda: Miền dựa trên mô men bậc cao
căn chỉnh cho việc thích ứng miền không giám sát. Hệ thống dựa
trên kiến thức, 261:110205, 2023.

[9] Đặng Kiến Khang, Giả Quốc, Lưu Thông Lương, Minh Minh Cung,
và Stefanos Zafeiriou. Mặt cung phụ ở giữa: Mặt tăng cường
nhận dạng bằng khuôn mặt web nhiễu quy mô lớn. Trong máy tính
Tầm nhìn–ECCV 2020: Hội nghị Châu Âu lần thứ 16, Glasgow,
Vương quốc Anh, ngày 23–28 tháng 8 năm 2020, Biên bản, Phần XI 16, trang
741–757. Springer, 2020.

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kot-sia và
Stefanos Zafeiriou. Retinaface: Định vị khuôn mặt nhiều cấp độ
một lần chụp trong tự nhiên. Trong Biên bản báo cáo
hội nghị IEEE/CVF về thị giác máy tính và mẫu
nhận dạng, trang 5203–5212, 2020.

[11] Đặng Kiến Khang, Giả Quả, Niannan Xue, và Stefanos
Zafeiriou. Arcface: Mất biên góc cộng cho sâu
nhận dạng khuôn mặt. Trong Biên bản hội nghị IEEE/CVF về thị giác
máy tính và nhận dạng mẫu, các trang
4690–4699, 2019.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov,
Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,
Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain
Gelly, et al. Một hình ảnh có giá trị bằng 16x16 từ: Trans-
formers để nhận dạng hình ảnh ở quy mô lớn. Bản in trước arXiv
arXiv:2010.11929, 2020.

[13] Yueqi Duan, Jiwen Lu và Jie Zhou. Uniformface: Học biểu diễn phân
phối đều sâu để nhận dạng khuôn mặt.
Trong Biên bản Hội nghị IEEE/CVF về Máy tính
Tầm nhìn và Nhận dạng Mẫu, trang 3415–3424, 2019.

[14] Yanbo Fan, Siwei Lyu, Yiming Ying và Baogang Hu.
Học tập với tổn thất top-k trung bình. Tiến bộ trong hệ thống xử
lý thông tin thần kinh, 30, 2017.

[15] Yuxin Fang, Ben Cheng Liao, Xinggang Wang, Jiemin Fang,
Jiyang Qi, Rui Wu, Jianwei Niu và Wenyu Liu. Bạn
chỉ nhìn vào một chuỗi: Xem xét lại máy biến áp trong tầm nhìn
thông qua phát hiện đối tượng. Những tiến bộ trong thông tin thần kinh
Hệ thống xử lý, 34:26183–26197, 2021.

[16] Mohsen Fayyaz, Soroush Abbasi Koohpayegani,
Farnoush Rezaei Jafari, Sunando Sengupta, Hamid
Reza Vaęzi Joze, Eric Sommerlade, Hamed Pirsiavash,
và Jurgen Gall. Lấy mẫu mã thông báo thích ứng cho các bộ chuyển
đổi thị giác hiệu quả. Trong Computer Vision–ECCV 2022: 17
Hội nghị Châu Âu, Tel Aviv, Israel, ngày 23–27 tháng 10,
2022, Biên bản báo cáo, Phần XI, trang 396–414. Springer, 2022.

[17] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian,
Chân Quân Hàn, Chu Bolei và Qixiang Ye. Ts-cam: Mã thông báo

bản đồ chú ý ngữ nghĩa kết hợp cho đối tượng được giám sát yếu
bản địa hóa. Trong Biên bản báo cáo của IEEE/CVF International
Hội nghị về Thị giác máy tính, trang 2886–2895, 2021.

[18] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, và
Qiang Liu. Cải thiện quá trình đào tạo máy biến đổi thị giác bằng cách
ngăn chặn quá trình làm mịn quá mức. Bản in trước arXiv arXiv:2104.12753,
4(11), 2021.

[19] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, và
Qiang Liu. Máy biến áp thị giác với sự đa dạng của bản vá.
Bản in trước của arXiv arXiv:2104.12753, 2021.

[20] Mengran Gou, Octavia Camps và Mario Sznaier. mẹ:
Tính năng thời điểm trung bình để xác định lại người. Trong
Biên bản hội nghị quốc tế IEEE về hội thảo thị giác máy tính, trang
1294–1303, 2017.

[21] Jia Guo, Jiankang Deng, Alexandros Lattas và Stefanos
Zafeiriou. Phân phối lại mẫu và tính toán để phát hiện khuôn mặt
hiệu quả. Trong Hội nghị quốc tế về Biểu diễn học tập, 2022.

[22] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu,
và Yunhe Wang. Biến áp trong biến áp. Tiến bộ
trong Hệ thống xử lý thông tin thần kinh, 34:15908–15919,
2021.

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

[24] Jie Hu, Li Shen và Gang Sun. Các mạng lưới ép và kích thích. Trong
Hội nghị IEEE/CVF năm 2018 về Tầm nhìn máy tính
và Nhận dạng Mẫu, trang 7132–7141, 2018.

[25] Gary B Huang, Marwan Mattar, Tamara Berg và Eric
Learned-Miller. Khuôn mặt được dán nhãn trong tự nhiên: Một 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 'đời thực': phát hiện,
căn chỉnh và nhận dạng, 2008.

[26] Yuge Huang, Pengchen Shen, Ying Tai, Shaoxin Li, Xi-aoming Liu,
Jilin Li, Feiyue Huang và Rongrong Ji. Cải thiện khả năng nhận
dạng khuôn mặt từ các mẫu cứng thông qua phân phối
tổn thất chưng cất. Trong Computer Vision–ECCV 2020: Hội nghị Châu
Âu lần thứ 16, Glasgow, Vương quốc Anh, 23–28 tháng 8 năm 2020,
Biên bản, Phần XXX 16, trang 138–154. Springer, 2020.

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

[28] Edwin T Jaynes. Lý thuyết thông tin và cơ học thống kê. Tạp chí vật
lý, 106(4):620, 1957.

[29] Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Jin-hua Ma, Yaowei
Wang và Wei-Shi Zheng. Chất làm giãn nở:
Máy biến áp giãn nở đa thang để nhận dạng trực quan. IEEE
Giao dịch về đa phương tiện, 2023.

[30] Kyungmin Kim, Bichen Wu, Xiaoliang Dai, Peizhao Zhang,
Zhicheng Yan, Peter Vajda và Seon Joo Kim. Suy nghĩ lại
sự tự chú ý trong các máy biến áp thị giác. Trong Biên bản
Hội nghị IEEE/CVF về Tầm nhìn máy tính và Mẫu
Nhận dạng, trang 3071–3075, 2021.

[31] Minchul Kim, Anil K Jain và Xiaoming Liu. Adaface:
Biên độ thích ứng chất lượng cho nhận dạng khuôn mặt. Trong Biên
bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và
Nhận dạng mẫu, trang 18750–18759, 2022.

[32] Alex Krizhevsky, Ilya Sutskever và Geoffrey E Hinton.
Phân loại Imagenet với mạng nơ-ron tích chập sâu. Truyền thông của
ACM, 60(6):84–90, 2017.

[33] Solomon Kullback. Lý thuyết thông tin và thống kê.
Công ty chuyển phát nhanh, 1997.

[34] Kunchang Li, Yali Wang, Gao Peng, Tanglu Song, Yu Liu,
Hongsheng Li và Yu Qiao. Uniformer: Máy biến áp thống nhất
để học biểu diễn không gian-thời gian hiệu quả. Trong Hội nghị quốc
tế về Biểu diễn học tập, 2021.

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

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, và
Piotr Dollar. Mất tiêu cự để phát hiện vật thể dày đặc. Trong Biên bản báo cáo
của hội nghị quốc tế IEEE về máy tính
tầm nhìn, trang 2980–2988, 2017.

[37] Hao Liu, Xiangyu Zhu, Zhen Lei và Stan Z Li. Khuôn mặt thích ứng:
Biên độ thích ứng và lấy mẫu để nhận dạng khuôn mặt. Trong
Biên bản Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu,
trang 11947–11956, 2019.

[38] Jihao Liu, Boxiao Liu, Hang Chu, Hongsheng Li và Yu
Liu. Tokenmix: Xem xét lại việc trộn hình ảnh để tăng cường dữ
liệu trong bộ chuyển đổi thị giác. Trong Computer Vision–ECCV
2022: Hội nghị Châu Âu lần thứ 17, Tel Aviv, Israel, 23–27 tháng
10 năm 2022, Biên bản báo cáo, Phần XXVI, trang 455–471.
Springer, 2022.

[39] Vị Ương Lưu, Văn Yên Đông, Trí Định Ngọc, Minh Lý, Tỳ Kheo
Raj và Le Song. Sphereface: Nhúng siêu cầu sâu
để nhận dạng khuôn mặt. Trong Biên bản báo cáo của hội nghị IEEE
về tầm nhìn máy tính và nhận dạng mẫu, trang 212–220,
2017.

[40] Dương Lưu, Đặng Kiến Khang, Phi Vương, Lôi Thượng, Huyền Tông
Xie và Baigui Sun. Damofd: Đào sâu vào thiết kế xương sống
về phát hiện khuôn mặt. Trong Hội nghị quốc tế lần thứ mười một
về Biểu diễn học tập, 2022.

[41] Dương Lưu và Từ Đường. Bfbox: Tìm kiếm khuôn mặt phù hợp
xương sống và mạng lưới kim tự tháp đặc trưng cho máy dò khuôn mặt. Trong
Biên bản Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu,
trang 13568–13577, 2020.

[42] Yang Liu, Xu Tang, Junyu Han, Jingtuo Liu, Dinger Rui,
và Xiang Wu. Hambox: Đi sâu vào khai thác chất lượng cao
neo về phát hiện khuôn mặt. Trong Hội nghị IEEE/CVF năm 2020
trên Computer Vision and Pattern Recognition (CVPR), các trang
13043–13051. IEEE, 2020.

[43] Yang Liu, Fei Wang, Jiankang Đặng, Zhipeng Chu, Baigui
Sun, và Hao Li. Mogface: Hướng tới sự đánh giá sâu sắc hơn
về phát hiện khuôn mặt. Trong Biên bản báo cáo của Hội nghị IEEE/
CVF về Thị giác máy tính và Nhận dạng mẫu, các trang
4093–4102, 2022.

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng
Zhang, Stephen Lin và Baining Guo. Máy biến áp Swin:

Bộ chuyển đổi tầm nhìn phân cấp sử dụng cửa sổ dịch chuyển. Trong Biên bản hội nghị quốc tế IEEE/CVF về tầm nhìn máy tính, trang 10012–10022, 2021.

[45] Ilya Loshchilov và Frank Hutter. Giảm trọng lượng tách rời chuẩn hóa. bản in trước arXiv arXiv:1711.05101, 2017.

[46] 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.

[47] Qiang Meng, Shichao Zhao, Zhida Huang và Feng Chu. Magface: Một đại diện phổ quát cho nhận dạng khuôn mặt và đánh giá chất lượng. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu, các trang 14225–14234, 2021.

[48] Stylianos Moschoglou, Athanasios Papaioannou, Chris-tos 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 của hội nghị IEEE về hội thảo về thị giác máy tính và nhận dạng mẫu, các trang 51–59, 2017.

[49] A Oppenheim, Jae Lim, Gary Kopec và SC Pohlig. Giai đoạn trong lời nói và hình ảnh. Trong ICASSP'79. IEEE International Hội nghị về Âm học, Giọng nói và Xử lý tín hiệu, tập 4, trang 632–637. IEEE, 1979.

[50] Alan V Oppenheim và Jae S Lim. Tầm quan trọng của pha trong tín hiệu. Biên bản báo cáo của IEEE, 69(5):529–541, 1981.

[51] Nikhil R Pal và Sankar K Pal. Entropy: Một định nghĩa mới và các ứng dụng của nó. Các giao dịch IEEE về hệ thống, con người và điều khiển học, 21(5):1260–1270, 1991.

[52] Leon N Piotrowski và Fergus W Campbell. Một minh chứng về tầm quan trọng và tính linh hoạt của biên độ và pha tần số không gian về mặt thị giác. Nhận thức, 11(3):337–346, 1982.

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

[54] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa 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.

[55] Claude Elwood Shannon. Một lý thuyết toán học về truyền thông. ACM SIGMOBILE mobile computing and communications review, 5(1):3–55, 2001.

[56] Abhinav Shrivastava, Abhinav Gupta và Ross Girshick. Đào tạo các máy dò đối tượng dựa trên vùng với khai thác ví dụ cứng trực tuyến. Trong Biên bản báo cáo của hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, trang 761–769, 2016.

[57] Kihyuk Sohn. Cải thiện việc học số liệu sâu với mục tiêu mất mát n-cặp đa lớp. Những tiến bộ trong thông tin thần kinh hệ thống xử lý, 29, 2016.

[58] Robin Strudel, Ricardo Garcia, Ivan Laptev và Cordelia Schmid. Segmenter: Bộ chuyển đổi cho phân đoạn ngữ nghĩa tion. Trong Biên bản báo cáo hội nghị quốc tế IEEE/CVF về thị giác máy tính, trang 7262–7272, 2021.

[59] Tôn Yifan, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang và Yichen Wei. Mất vòng tròn: Một góc nhìn 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 nhận dạng mẫu, trang 6398–6407, 2020.

[60] Tôn Chân Hồng, Minh Lâm, Tôn Tú Ngọc, Tân Trí Ngọc, Hạo Li, và Rong Jin. Mae-det: Xem xét lại nguyên lý entropy tối đa trong nas zero-shot để phát hiện đối tượng hiệu quả. Trong Hội nghị quốc tế về Học máy, trang 20810–20826. PMLR, 2022.

[61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles và Herve J egou. Đào tạo bộ chuyển đổi hình ảnh hiệu quả về dữ liệu & chưng cất thông qua sự chú ý. Trong hội nghị quốc tế về học máy, trang 10347–10357. PMLR, 2021.

[62] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve và Herve J egou. Đi sâu hơn với bộ biến đổi hình ảnh. Trong Biên bản báo cáo Hội nghị quốc tế về thị giác máy tính của IEEE/CVF, trang 32–42, 2021.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser và Illia Polosukhin. Sự chú ý là tất cả những gì bạn cần. Những tiến bộ trong thần kinh hệ thống xử lý thông tin, 30, 2017.

[64] Feng Wang, Jian Cheng, Weiyang Liu, và Haijun Liu. Biên độ cộng softmax để xác minh khuôn mặt. IEEE Signal Pro-processing Letters, 25(7):926–930, 2018.

[65] Hao Wang, Yitong Wang, Zheng Chu, Xing Ji, Dihong Gong, Jingchao Chu, Zhifeng Li và Wei Liu. Hóa trang: Tổn thất 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.

[66] Xiaobo Wang, Shuo Wang, Shifeng Zhang, Tianyu Fu, Hailin Shi và Tao Mei. Hỗ trợ vector hướng dẫn softmax mất mát cho nhận dạng khuôn mặt. bản in trước arXiv arXiv:1812.11317, 2018.

[67] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi và Tao Mei. Vector phân loại sai hướng dẫn mất mát soft-max cho nhận dạng khuôn mặt. Trong Biên bản của AAAI Hội nghị về Trí tuệ nhân tạo, trang 12241–12248, 2020.

[68] Yandong Wen, Kaipeng Zhang, Zhifeng Li, và Yu Qiao. MỘT phương pháp học đặc điểm phân biệt cho nhận dạng khuôn mặt sâu. Trong Computer Vision–ECCV 2016: Hội nghị châu Âu lần thứ 14 Hội nghị, Amsterdam, Hà Lan, ngày 11–14 tháng 10, 2016, Biên bản, Phần VII 14, trang 499–515. Springer, 2016.

[69] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollar và Ross Girshick. Các phép tích chập ban đầu giúp người chuyển đổi nhìn rõ hơn. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 34:30392–30400, 2021.

[70] Mingle Xu, Sook Yoon, Alvaro Fuentes và Dong Sun Park. Một cuộc khảo sát toàn diện về các kỹ thuật tăng cường hình ảnh để học sâu. Nhận dạng mẫu, trang 109347, 2023.

[71] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, và Qi Tian. Một khuôn khổ dựa trên fourier để tổng quát hóa miền

tion. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu, trang 14383–14392, 2021.

[72] Weijian Xu, Yifan Xu, Tyler Chang và Zhuowen Tu. Bộ chuyển đổi hình ảnh chuyển đổi chú ý theo tỷ lệ đồng quy. Trong Biên bản báo cáo của Hội nghị quốc tế IEEE/CVF về thị giác máy tính, trang 9981–9990, 2021.

[73] Yanchao Yang và Stefano Soatto. Fda: Miền Fourier sự thích nghi cho phân đoạn ngữ nghĩa. Trong Biên bản Hội nghị IEEE/CVF về Tầm nhìn máy tính và Mẫu Nhận dạng, trang 4085–4095, 2020.

[74] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Chu, Feng-wei Yu và Wei Wu. Kết hợp các thiết kế tích chập vào máy biến áp trực quan. Trong Biên bản báo cáo của Hội nghị quốc tế về Tầm nhìn máy tính IEEE/CVF, trang 579–588, 2021.

[75] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe và Youngjoon Yoo. Cutmix: Chiến lược chuẩn hóa để đào tạo các bộ phân loại mạnh với khả năng bản địa hóa tính năng. Trong Biên bản báo cáo hội nghị quốc tế IEEE/CVF về thị giác máy tính, trang 6023–6032, 2019.

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

[77] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang và Hong-sheng Li. Adacos: Điều chỉnh tỷ lệ logit cosine một cách thích ứng để học hiệu quả các biểu diễn khuôn mặt sâu. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Tầm nhìn máy tính và Mẫu Công nhận, trang 10823–10832, 2019.

[78] Tiểu Trương, Rui Zhao, Junjie Yan, Mengya Gao, Yu Qiao, Xiaogang Wang và Hongsheng Li. P2sgrad: Các gra-dient tinh chỉnh để tối ưu hóa các mô hình mặt sâu. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Tầm nhìn máy tính và Mẫu Công nhận, trang 9906–9914, 2019.

[79] Kai Zhao, Jingyi Xu, và Ming-Ming Cheng. Khuôn mặt thông thường: Nhận dạng khuôn mặt sâu thông qua chính quy hóa độc quyền. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng Mẫu, trang 1136–1144, 2019.

[80] Yaoyao Zhong và Weihong Đặng. Máy biến áp mặt cho sự công nhận. bản in trước arXiv arXiv:2103.14803, 2021.

[81] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, và Yi Yang. Tăng cường dữ liệu xóa ngẫu nhiên. Trong Biên bản báo cáo của hội nghị AAAI về trí tuệ nhân tạo, các trang 13001–13008, 2020.

[82] Benjia Zhou, Picao Wang, Jun Wan, Yanyan Liang và Fan Wang. Đào tạo chuyển đổi tầm nhìn hiệu quả: Một dữ liệu tập trung quan điểm. bản in trước arXiv arXiv:2209.15006, 2022.