**Khái niệm Face Recognition:**

* Hệ thống nhận dạng khuôn mặt là một ứng dụng máy tính
* Tự động xác định hoặc nhận dạng một người nào đó từ một bức hình ảnh kỹ thuật số hoặc một khung hình video từ một nguồn video.
* Một trong những cách để thực hiện điều này là so sánh các đặc điểm khuôn mặt chọn trước từ hình ảnh và một cơ sở dữ liệu về khuôn mặt

**Ứng dụng của FR:**

1. Mở khóa điện thoại thông minh

2. Giám sát và an ninh tại sân bay

3. Tự động gắn thẻ trong mạng xã hội

4. Xác thực danh tính tại máy ATM

5. Quản lý ra vào tại văn phòng làm việc

6. Kiểm tra danh tính hành khách trong giao thông công cộng

,,,

**Hệ thống nhận diện khuôn mặt thường bao gồm các thành phần chính sau:**

* **Cảm biến hình ảnh**: Đây là phần cứng, thường là camera, dùng để thu thập hình ảnh khuôn mặt.
* **Tiền xử lý hình ảnh**: Giai đoạn này giúp cải thiện chất lượng hình ảnh, bao gồm việc làm sạch, tăng cường độ tương phản, và điều chỉnh ánh sáng.
* **Phát hiện khuôn mặt**: Sử dụng các thuật toán để xác định vị trí khuôn mặt trong hình ảnh, loại bỏ các phần không cần thiết.
* **Trích xuất đặc trưng**: Giai đoạn này chuyển đổi hình ảnh khuôn mặt thành một dạng số để dễ dàng phân tích, thường bằng cách sử dụng các phương pháp như Deep Learning hoặc Eigenfaces.
* **Nhận diện và so khớp:** So sánh các đặc trưng đã trích xuất với cơ sở dữ liệu đã lưu để xác định danh tính.
* *Quản lý dữ liệu: Lưu trữ và quản lý thông tin về các khuôn mặt đã được nhận diện.*
* *Giao diện người dùng: Cung cấp cách thức tương tác với hệ thống, giúp người dùng có thể nhập dữ liệu hoặc nhận kết quả.*
* *Hệ thống phản hồi: Cung cấp thông tin cho người dùng về kết quả nhận diện, bao gồm xác suất và độ tin cậy của kết quả.*

**Những thách thức đặt ra với Face Recognition**

**Độ chính xác và độ tin cậy**: Hệ thống cần đạt độ chính xác cao trong nhiều điều kiện khác nhau, nhưng có thể gặp khó khăn trong việc nhận diện đúng người khi có ánh sáng yếu, góc nhìn khác nhau hoặc khi người dùng đeo kính, khẩu trang.

**Đối tượng đa dạng**: Các khuôn mặt đến từ nhiều nền văn hóa và sắc tộc khác nhau có thể gây ra vấn đề trong việc phân loại và nhận diện.

* **Dữ liệu huấn luyện đa dạng**: Tăng cường tập dữ liệu huấn luyện bằng cách thu thập hình ảnh khuôn mặt từ nhiều nền văn hóa, sắc tộc, và độ tuổi khác nhau. Điều này giúp hệ thống học cách nhận diện và phân loại các khuôn mặt đa dạng hơn.
* **Kỹ thuật chuyển giao học**: Sử dụng các mô hình đã được huấn luyện trên dữ liệu lớn và đa dạng, sau đó tinh chỉnh chúng cho các tập dữ liệu cụ thể, giúp cải thiện khả năng nhận diện cho những nhóm chưa được đại diện đầy đủ.
* **Kiểm tra và đánh giá công bằng**: Thực hiện các bài kiểm tra định kỳ để đánh giá hiệu suất của hệ thống trên các nhóm người khác nhau, nhằm phát hiện và khắc phục sự thiên lệch.

**Thời gian xử lý**: Để đảm bảo hiệu quả trong các ứng dụng thực tế, thời gian nhận diện cần phải nhanh chóng, nhưng các thuật toán phức tạp có thể làm chậm quá trình.

**Bảo mật và quyền riêng tư**: Đảm bảo an toàn cho dữ liệu cá nhân và tuân thủ các quy định về bảo vệ dữ liệu.

* **Chính sách bảo mật rõ ràng**: Cần có chính sách minh bạch về cách thức thu thập, sử dụng, và lưu trữ dữ liệu nhận diện khuôn mặt. Người dùng nên được thông báo rõ ràng về cách thông tin của họ sẽ được sử dụng.
* **Đồng ý của người dùng**: Trước khi thu thập dữ liệu khuôn mặt, cần có sự đồng ý rõ ràng từ người dùng. Hệ thống nên cung cấp tùy chọn cho người dùng để đồng ý hoặc từ chối.
* **Ẩn danh dữ liệu**: Sử dụng các kỹ thuật ẩn danh để đảm bảo rằng dữ liệu thu thập không thể dễ dàng liên kết với danh tính cá nhân. Điều này có thể bao gồm mã hóa hoặc làm mờ dữ liệu khuôn mặt.
* **Giới hạn lưu trữ dữ liệu**: Thiết lập thời gian lưu trữ tối đa cho dữ liệu nhận diện khuôn mặt và tự động xóa dữ liệu khi không còn cần thiết.
* **Bảo mật dữ liệu**: Áp dụng các biện pháp bảo mật mạnh mẽ để bảo vệ dữ liệu khỏi truy cập trái phép, bao gồm mã hóa và kiểm soát truy cập.

Vision transformer

### Khái niệm: Vision Transformer (ViT) là gì?

### Vision Transformer (ViT) là gì?

ViT là một mô hình trực quan **dựa trên kiến ​​trúc của một kiến trúc Transformer** ban đầu được thiết kế cho các tác vụ dựa trên văn bản. **Mô hình ViT chuyển hình ảnh đầu vào thành một loạt các mảng hình ảnh**, và **dự đoán trực tiếp các nhãn lớp cho hình ảnh**. ViT cho thấy một hiệu suất phi thường khi được huấn luyện trên đủ dữ liệu, vượt xa hiệu suất của một CNN hiện đại tương tự với tài nguyên tính toán ít hơn 4 lần.

1. Chia hình ảnh thành các mảng (patch) với kích thước từng mảng cố định
2. Làm phẳng các mảng hình ảnh
3. Tạo các feature embedding có chiều thấp hơn từ các mảng hình ảnh phẳng này
4. Bao gồm thứ tự các mảng
5. Chuỗi feature embedding được làm đầu vào cho transformer encoder
6. Thực hiện pre-train đối với mô hình ViT với các nhãn hình ảnh, sau đó được giám sát hoàn toàn trên một tập dữ liệu lớn
7. Tinh chỉnh model trên bộ dữ liệu riêng của từng bài toán

Mô hình ViT sử dụng Multi-Head Self Attention trong lĩnh vực Thị giác máy. Mô hình này chia hình ảnh lớn ra thành một loạt các mảng hình ảnh nhỏ hơn và được nén lại thành vector theo vị trí của ảnh, sau đó được đưa vào Transformer encoder. Khi làm như vậy, Transformer encoder học được những đặc tính tổng quát và chi tiết có trong ảnh. Cuối cùng, ViT không những có độ chính xác cao hơn trên những dataset lớn, mà còn giảm lượng thời gian cần thiết để huấn luyện mô hình.

Ưu điểm của ViT

**Khả năng học mối quan hệ dài hạn** (Long-range dependencies)

Transformer sử dụng cơ chế tự chú ý (self-attention), cho phép mô hình học được mối quan hệ giữa các phần khác nhau của hình ảnh một cách trực tiếp mà không bị giới hạn bởi khoảng cách không gian như trong CNN.

Điều này giúp ViT nhận biết được các kết nối giữa các vùng khác nhau trên ảnh, điều mà CNN có thể bỏ sót do tính cục bộ của convolution.

**Không phụ thuộc vào cấu trúc không gian cục bộ**

CNN dựa trên các phép convolution cục bộ, thường yêu cầu một kiến trúc cố định về kích thước bộ lọc và cấu trúc của từng lớp. Ngược lại, Vision Transformer không cần biết về cấu trúc không gian của hình ảnh mà vẫn có thể xử lý tốt các thông tin này qua cách phân tích các patch (mảnh nhỏ) của ảnh.

Điều này cho phép ViT tổng quát hóa tốt hơn trên các tác vụ và có khả năng mở rộng dễ dàng.

**Tính tổng quát hóa mạnh mẽ:** Vision Transformer thường yêu cầu một lượng dữ liệu lớn để huấn luyện và có khả năng tổng quát hóa tốt hơn trên nhiều tập dữ liệu. Điều này giúp ViT hoạt động hiệu quả trong nhiều lĩnh vực khác nhau, đặc biệt là khi có sẵn các mô hình pre-trained lớn.

Khi được huấn luyện trên các bộ dữ liệu lớn (như ImageNet), ViT có thể đạt được kết quả tốt hơn so với CNN trong nhiều bài toán thị giác máy tính.

Transformer trong Face Recognition

1. Yaoyao Zhong, Weihong Deng :"Face Transformer for Recognition", 2021

2. Zhonglin Sun :"Part-based Face Recognition with Vision Transformers"

3. Jun Dan, Yang Liu :"TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective"

4. Niloufar Alipour Talemi :"CATFace: Cross-Attribute-Guided Transformer with Self-Attention Distillation for Low-Quality Face Recognition"

5. Minchul Kim :"KeyPoint Relative Position Encoding for Face Recognition"

--> scan qua các bài này, survey thêm 5 bài nữa

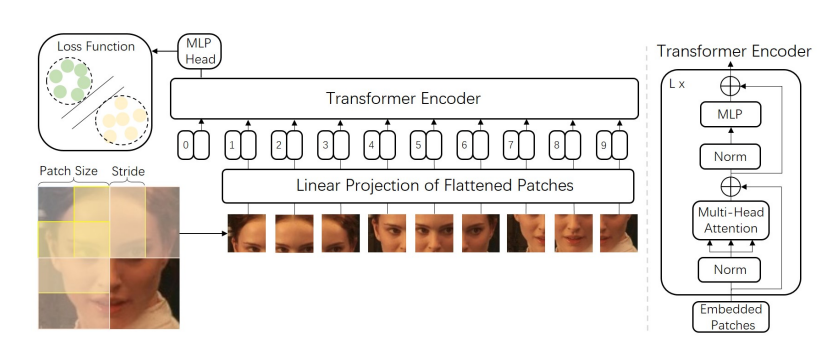
--> out:

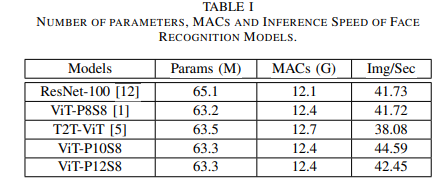
việc TS ứng dụng trong FR có tốt ko , họ đánh giá nó ntn

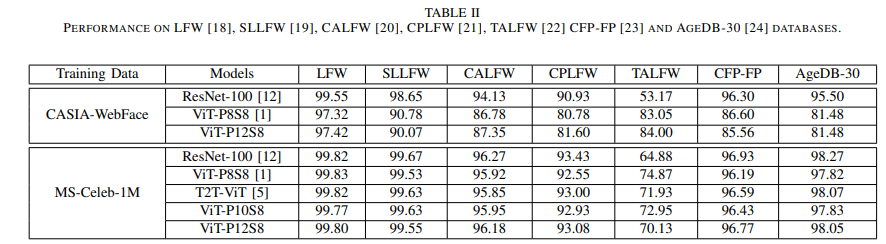
Tính khả thi, những ưu điểm nhược điểm

Bài 1: Face Transformer for Recognition (Yaoyao Zhong, Weihong Deng)

* **Hiệu suất:** Trên tập dữ liệu lớn, mô hình Transformer có hiệu suất cạnh tranh với các mô hình CNN như ResNet, đặc biệt khi sử dụng kỹ thuật chồng lấn patch
* **khả năng chịu đựng các tác nhân gây nhiễu không tốt** bằng CNN trong việc xử lý hình ảnh bị che khuất







Nội dung chính của bài báo:

Mục tiêu nghiên cứu: Bài báo nhằm khám phá khả năng của các mô hình Transformer trong nhận diện khuôn mặt, đặc biệt là so với CNN, vốn đã rất thành công trong lĩnh vực này.

Phương pháp tiếp cận: Tác giả đã điều chỉnh quy trình tạo token để cải thiện khả năng mô tả thông tin giữa các patch (mảnh ảnh) bằng cách sử dụng các patch trượt có sự chồng lấp. Điều này giúp tăng cường khả năng nhận diện các đặc điểm khuôn mặt quan trọng mà có thể bị phân tán giữa các token khác nhau.

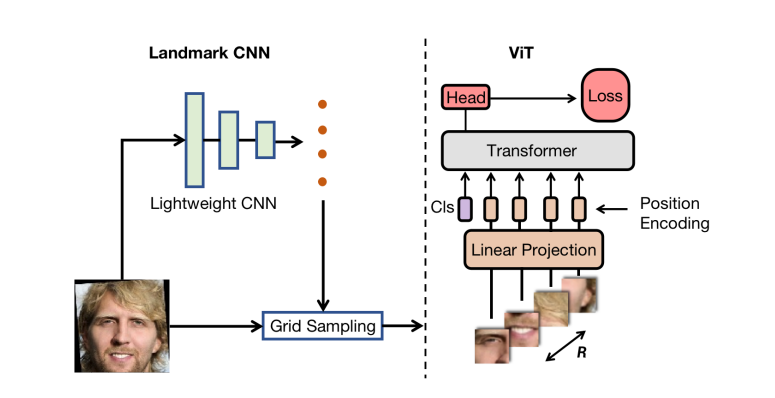
Dữ liệu và thử nghiệm: Các mô hình được đào tạo trên hai cơ sở dữ liệu lớn là CASIA-WebFace và MS-Celeb-1M, và được đánh giá trên nhiều benchmark phổ biến như LFW (Labeled Faces in the Wild), SLLFW, CALFW, CPLFW, TALFW, CFP-FP, AGEDB và IJB-C.

Kết quả: Kết quả cho thấy rằng các mô hình Face Transformer đạt được hiệu suất tương đương với CNN trong việc nhận diện khuôn mặt, mặc dù có một số khác biệt về khả năng tổng quát khi áp dụng trên các tập dữ liệu khác nhau.

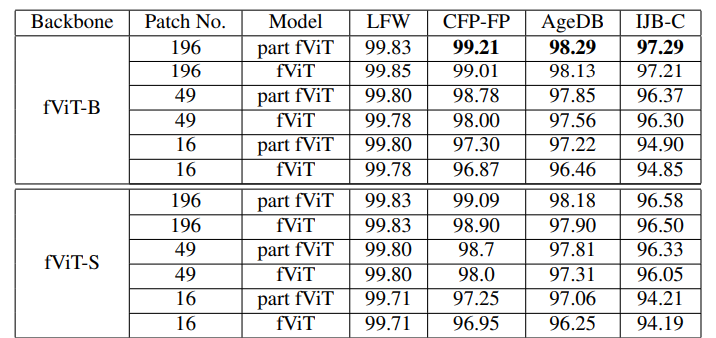
Kết luận: Bài báo khẳng định tính khả thi của việc áp dụng mô hình Transformer trong nhận diện khuôn mặt và mở ra hướng nghiên cứu mới để cải thiện hiệu suất và hiệu quả của các mô hình này trong tương lai.

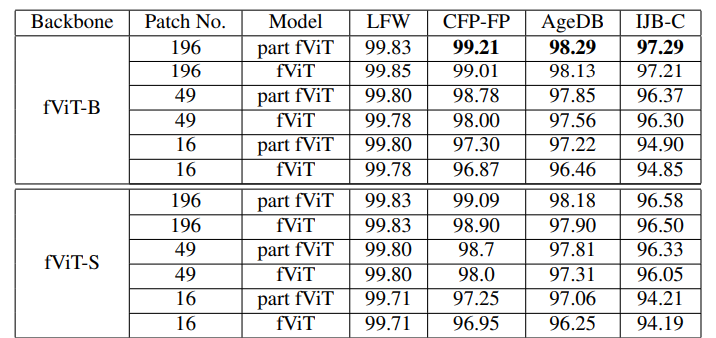
Bài 2:

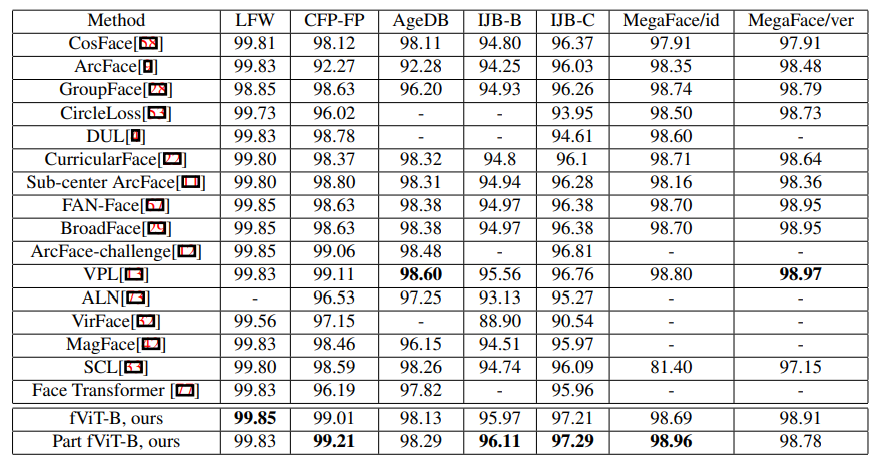
* Các phương pháp hiện có xử lý khuôn mặt một cách toàn diện dựa tên mạng nơ ron tích chập (CNN)
* Sự xuất hiện của ViT gần đây mang lại hiệu quả khá ấn tượng.
* Tác giả cho rằng “Xử lý khuôn mặt theo một cấu trúc nhiều phần (Mắt, mũi, miệng) có thể mang lại hiệu quả cao hơn”





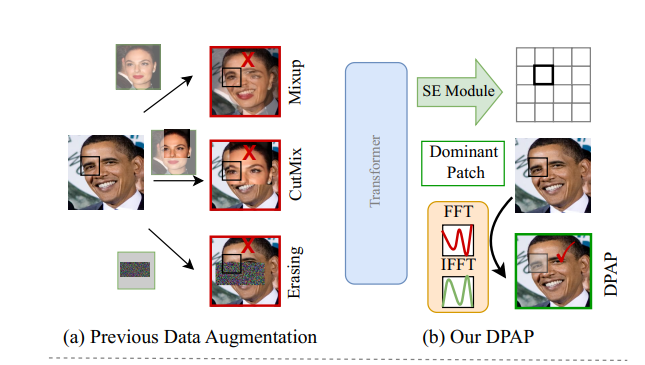






Bài 3:

* Các phương pháp hiện tăng cường dữ liệu hiện tại chưa phù hợp với mô hình Transformer nói chung và ViT nói riêng
* Xuất phát từ các chiến lược tăng cường DL chưa phù hợp với cấu trúc của ViTs.
* Overfiting và sự phụ thuộc quá mức vào một số đặc trưng khuôn mặt.



**Bài 3: Jun Dan, Yang Liu :"TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective"**

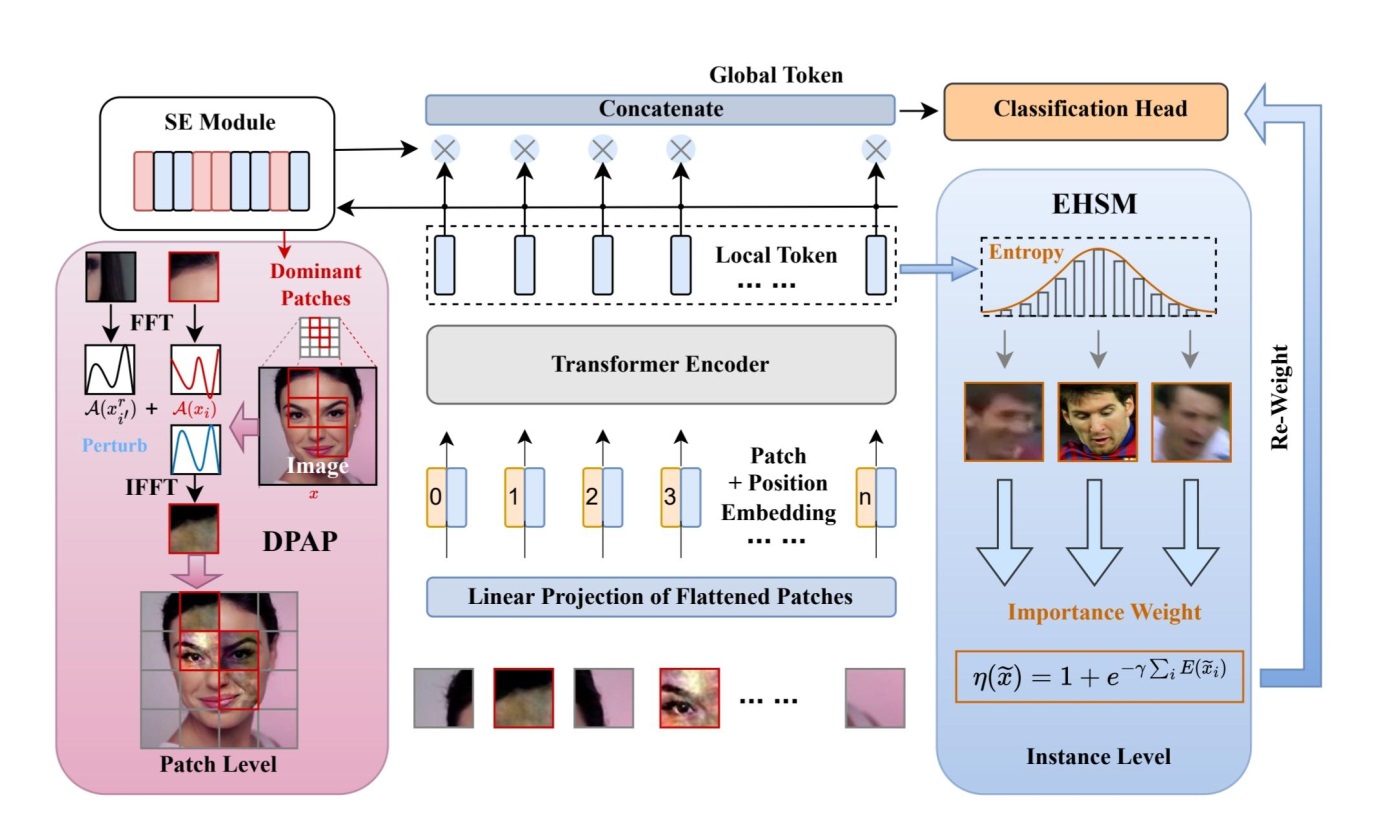
**Những vấn đề chính:**

* **ViTs và hạn chế trong Face Recognition (FR):**
  + Vision Transformers vốn dĩ yêu cầu dữ liệu rất lớn để phát huy khả năng, nhưng khi áp dụng vào bài toán nhận diện khuôn mặt với lượng dữ liệu lớn, hiệu suất của ViTs lại gần như tương đương với CNNs.
  + Việc sử dụng các chiến lược **data augmentation** thông thường như Random Erasing, Mixup, CutMix... dẫn đến việc làm mất đi các thông tin cấu trúc quan trọng của khuôn mặt, gây ảnh hưởng tiêu cực đến quá trình tối ưu hóa của ViTs. Ngoài ra, ViTs dễ bị **overfitting** khi huấn luyện trên các mảng khuôn mặt nhỏ lẻ (local patches), điều này gây suy giảm khả năng tổng quát hóa của mô hình.

### Giải pháp đề xuất:

1. **Patch-level Data Augmentation Strategy (DPAP):**
   * **DPAP** được thiết kế để tránh mất thông tin cấu trúc khuôn mặt khi thực hiện augmentation.
   * Phương pháp này sử dụng mô-đun **Squeeze-and-Excitation (SE)** để chọn ra các **dominant patches** (các mảng chính), sau đó trộn ngẫu nhiên thông tin biên độ của những mảng này mà không làm mất đi thông tin pha. Điều này giúp tạo ra các mẫu đa dạng hơn, tránh hiện tượng **overfitting**.
   * DPAP khuyến khích ViTs sử dụng thêm các phần khác của khuôn mặt, ví dụ như tai, mũi, miệng – những khu vực thường bị bỏ qua – giúp mô hình đưa ra quyết định tự tin hơn.
2. **Hard Sample Mining Strategy (EHSM):**
   * Chiến lược **Entropy-guided Hard Sample Mining (EHSM)** được đề xuất để khai thác tốt hơn các mẫu khó.
   * EHSM xem ViT như một hệ thống xử lý thông tin, sử dụng lý thuyết thông tin để đo lường lượng thông tin chứa trong các local tokens, từ đó điều chỉnh động trọng số của các mẫu dễ và khó trong quá trình huấn luyện. Điều này giúp mô hình khai thác triệt để các thông tin chi tiết trong từng khu vực của khuôn mặt (ví dụ: môi, cằm), tăng cường khả năng biểu diễn đặc trưng của từng token.

Cụ thể quá trình thực hiện



### Patch-Level Data Augmentation Strategy

Các công trình trước đây chủ yếu áp dụng các chiến lược **data augmentation** ở cấp độ mẫu để giảm thiểu hiện tượng **overfitting** trong ViTs. Tuy nhiên, các chiến lược này thường làm mất đi một số thông tin cấu trúc quan trọng của khuôn mặt, ảnh hưởng đến quá trình học các **face tokens** phân biệt.

Để giải quyết vấn đề này, nhóm tác giả đề xuất chiến lược **patch-level data augmentation** mới có tên **Dominant Patch Amplitude Perturbation (DPAP)**. Các bước chính của chiến lược này bao gồm:

1. Chèn một **SE module** (S) tại đầu ra của **transformer encoder** (F) và sử dụng các **scaling factors** do mô-đun S tạo ra để tìm ra các **dominant patches** quan trọng nhất.
2. Áp dụng cơ chế **linear mixing** để ngẫu nhiên làm nhiễu thông tin biên độ của các **dominant patches**.
3. Đưa ảnh đã được tái cấu trúc vào mô hình **TransFace** để huấn luyện.

Về mặt toán học, sau khi các **local tokens** được rút trích bởi **transformer encoder**, chúng được tái cân bằng qua mô-đun S. Điều này giúp mô hình không bị quá phụ thuộc vào các **dominant patches** và thay vào đó, khám phá thêm các vùng khác của khuôn mặt như tai, mũi, và miệng, nhằm đưa ra dự đoán tốt hơn.

### Entropy-based Hard Sample Mining Strategy

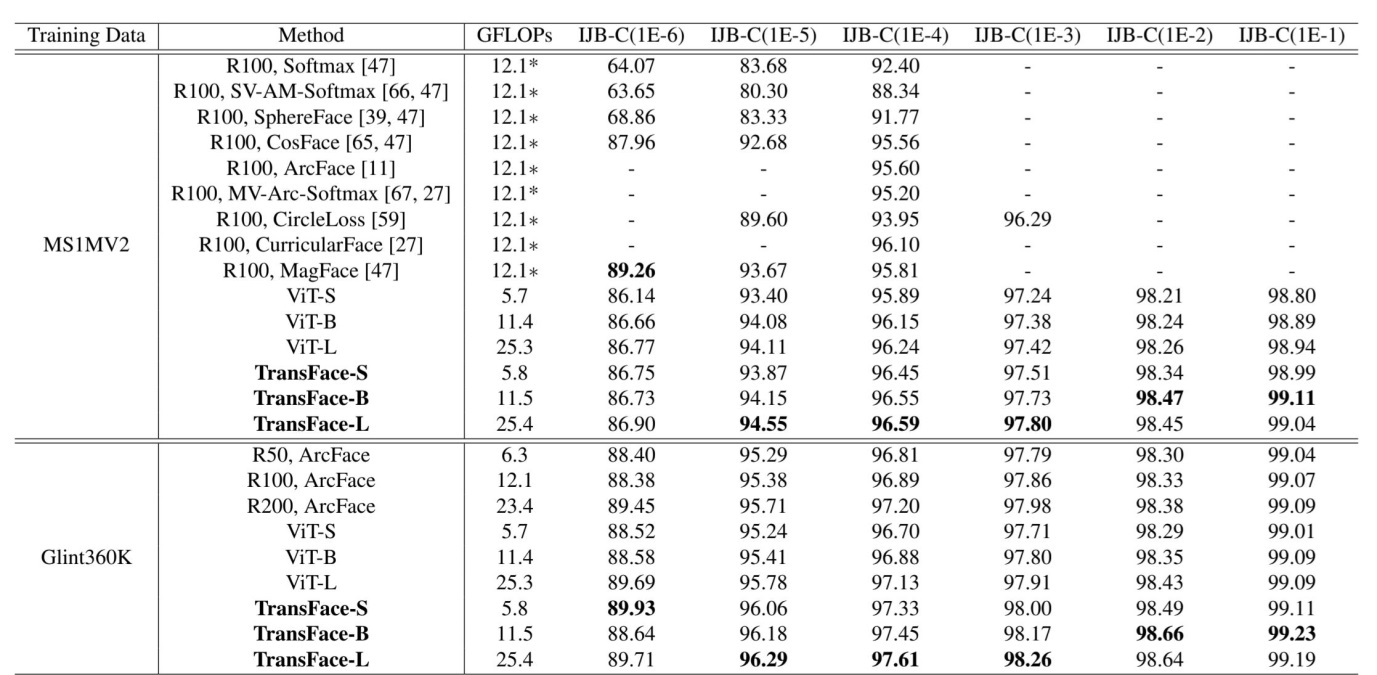
Công nghệ **hard sample mining** được sử dụng để tăng cường hiệu suất của mô hình bằng **cách khuyến khích mô hình học từ các mẫu khó hơn.** Các phương pháp khai thác mẫu khó trước đây chủ yếu sử dụng các chỉ số ở cấp độ mẫu, chẳng hạn như xác suất dự đoán hoặc **prediction loss**. Tuy nhiên, trong ViTs, việc dự đoán chủ yếu phụ thuộc vào một số ít **local tokens**, do đó việc sử dụng các chỉ số này có thể không tối ưu.

Để khai thác mẫu khó một cách chính xác hơn, nhóm tác giả đề xuất phương pháp **Entropy-guided Hard Sample Mining (EHSM)**. **EHSM** đo lường độ khó của mẫu dựa trên tổng lượng thông tin chứa trong các **local tokens**. Đối với các mẫu có nhiều thông tin (mẫu dễ), chúng có xu hướng chứa nhiều **information entropy** hơn và dễ học hơn. Ngược lại, các mẫu có ít thông tin (mẫu khó) sẽ có ít **information entropy** hơn, khiến chúng khó học hơn.

Cuối cùng, EHSM sử dụng một cơ chế trọng số dựa trên entropy để thích ứng với từng mẫu trong quá trình huấn luyện.

### Thử nghiệm và kết quả:

* Kết quả thực nghiệm trên các bộ dữ liệu nhận diện khuôn mặt nổi tiếng (như **IJB-C**, **Glint360K**) cho thấy mô hình **TransFace** vượt trội so với các mô hình ViTs và CNNs thông thường.
* Đặc biệt, mô hình đạt độ chính xác 97.61% trên chỉ số "TAR@FAR=1E-4" khi sử dụng bộ dữ liệu **Glint360K**

****

**KẾT LUẬN**

1. Đề xuất chiến lược **DPAP** giúp giải quyết vấn đề **overfitting** của ViTs trong bài toán nhận diện khuôn mặt.
2. Giới thiệu chiến lược **EHSM** để tăng tính ổn định trong việc dự đoán của mô hình nhận diện khuôn mặt.
3. Kết quả thử nghiệm chứng minh tính ưu việt của mô hình **TransFace** trên các bộ dữ liệu tiêu chuẩn.

Bài 4:

 **Vấn đề**: Các mô hình ViT thường gặp khó khăn khi xử lý các biến dạng hình học trong dữ liệu đầu vào, đặc biệt khi hình ảnh khuôn mặt không được căn chỉnh chính xác (alignment failures). Điều này xảy ra thường xuyên với các hình ảnh chất lượng thấp như ảnh từ camera giám sát, nơi việc nhận diện điểm mốc khuôn mặt có thể không chính xác.

 **Giải pháp**: Bài báo đề xuất một phương pháp mới gọi là **KP-RPE (KeyPoint Relative Position Encoding)**, sử dụng các điểm mốc trên khuôn mặt (như mắt, mũi, miệng) để tạo ra các vị trí tương đối giữa các điểm ảnh, giúp mô hình ViT duy trì mối quan hệ không gian giữa các bộ phận trên khuôn mặt dù có biến dạng hình học.

 **Lợi ích**: Phương pháp KP-RPE giúp mô hình ViT trở nên **mạnh mẽ hơn trước các biến dạng hình học** mà không cần căn chỉnh chính xác, nâng cao hiệu suất trong các tác vụ nhận diện khuôn mặt. Phương pháp này cũng hiệu quả đối với các hình ảnh không căn chỉnh tốt và dữ liệu chất lượng thấp như trong bộ dữ liệu **TinyFace** và **IJB-S**.

 **Kết quả thực nghiệm**: KP-RPE được chứng minh là cải thiện hiệu suất trên cả các tập dữ liệu hình ảnh căn chỉnh và không căn chỉnh, mà không gây tăng đáng kể về chi phí tính toán.