**TÓM TẮT NỘI DUNG KIẾN THỨC TUẦN 4**

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

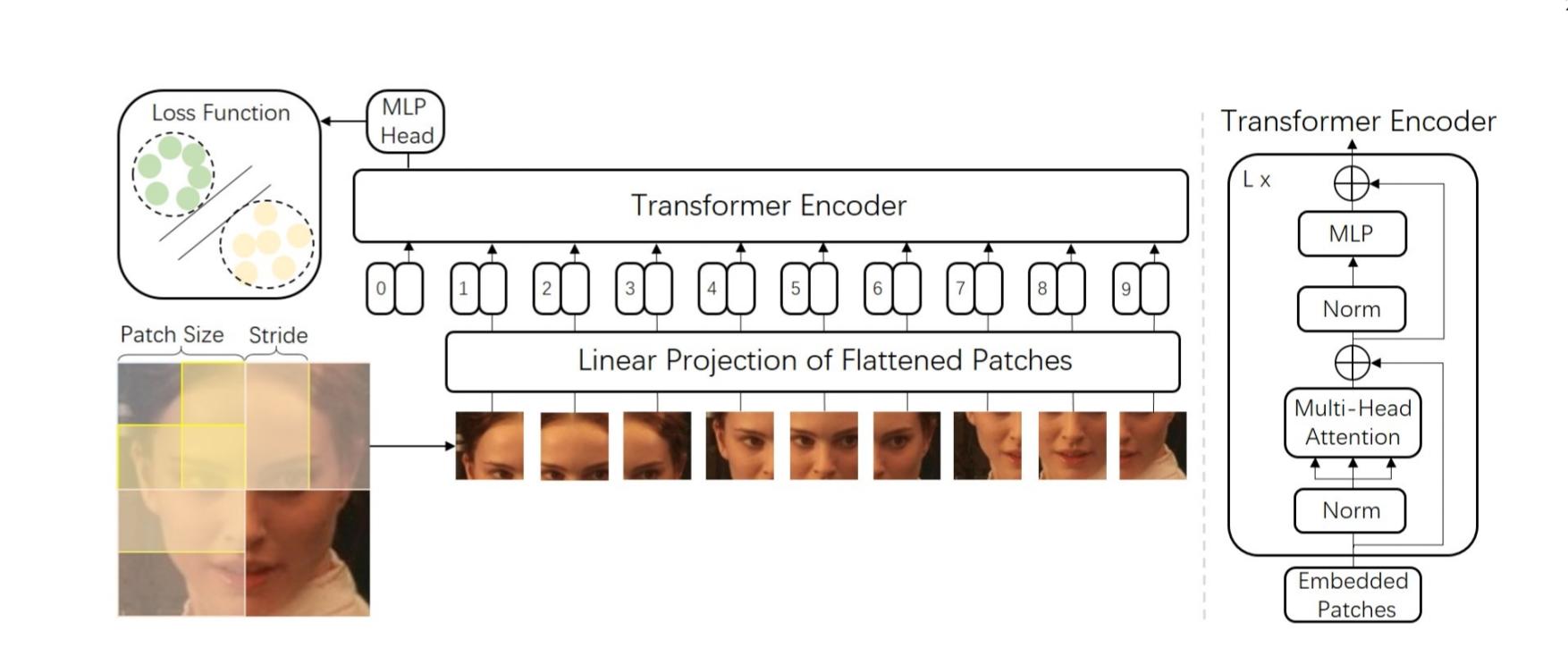
Tính khả thi của việc ứng dụng Transformer (TS) trong Nhận diện khuôn mặt (Face Recognition - FR):

Trong bài báo, nhóm tác giả đã chứng minh rằng mô hình Transformer có khả năng áp dụng được vào nhận diện khuôn mặt, nhưng tính khả thi phụ thuộc lớn vào kích thước của tập dữ liệu huấn luyện:

* **Tập dữ liệu nhỏ**: Khi huấn luyện trên tập dữ liệu nhỏ như **CASIA-WebFace**, Transformer không hoạt động tốt bằng CNN (cụ thể là ResNet-100). Điều này cho thấy mô hình Transformer cần một lượng lớn dữ liệu để phát huy hiệu quả.
* **Tập dữ liệu lớn**: Khi huấn luyện trên tập dữ liệu lớn như **MS-Celeb-1M**, Transformer đạt được hiệu suất gần tương đương với ResNet-100 về độ chính xác và số lượng tham số.

Kết luận của nhóm tác giả là mô hình Transformer có khả thi khi áp dụng cho nhận diện khuôn mặt, đặc biệt khi có dữ liệu lớn.

Kỹ thuật chồng lấn patch:



Ưu điểm của kỹ thuật chồng lấn patch:

* Nâng cao hiệu quả mô hình
* Không tăng đáng kể chi phí tính toán

**2. Đánh giá của nhóm tác giả về Transformer trong FR:**

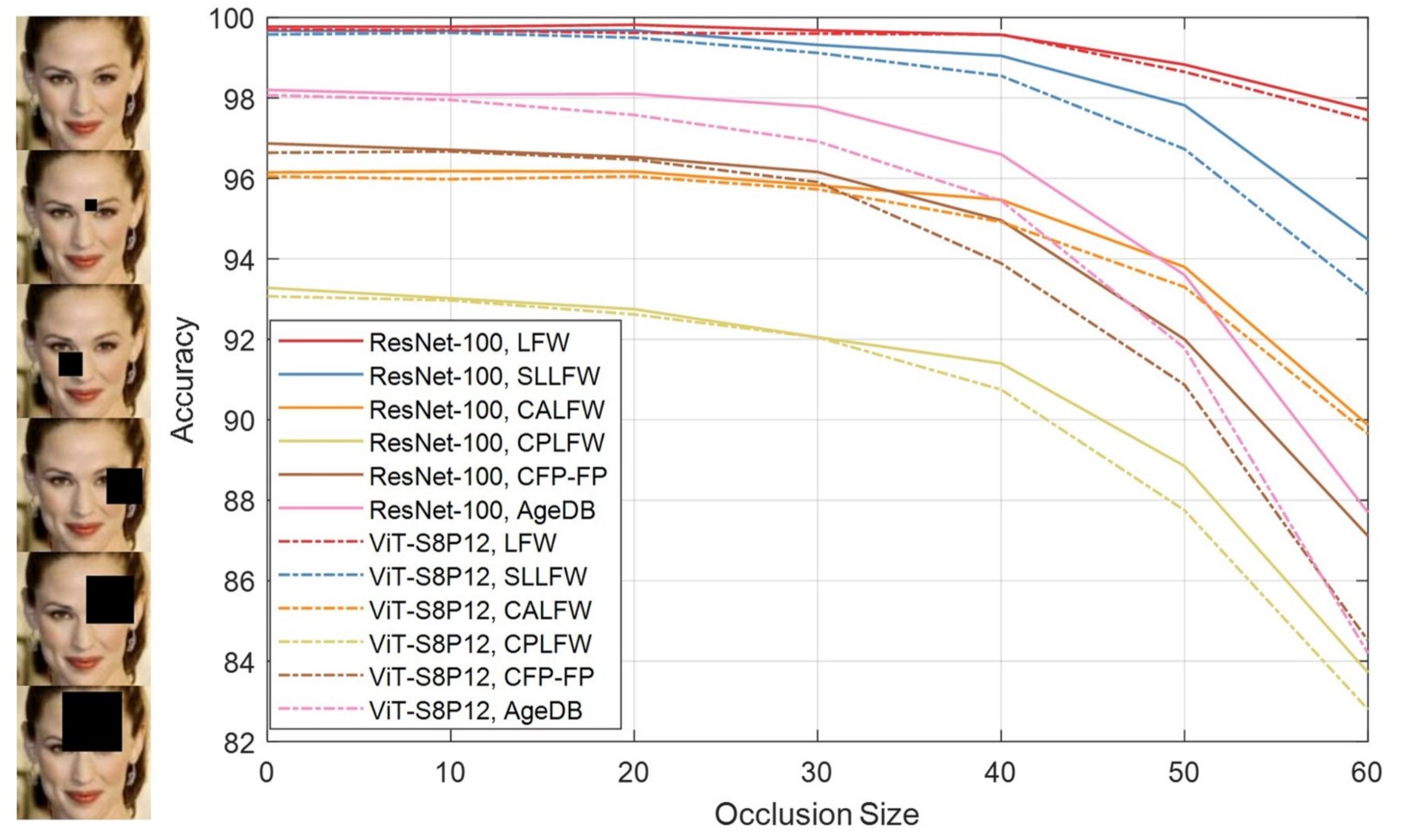
* **Hiệu suất:** Trên tập dữ liệu lớn, mô hình Transformer có hiệu suất cạnh tranh với các mô hình CNN như ResNet, đặc biệt khi sử dụng kỹ thuật chồng lấn patch (overlapping patches) để thu thập thông tin tốt hơn từ các vùng ảnh.
* **Robustness (khả năng chịu đựng các tác nhân gây nhiễu)**: Transformer hoạt động tốt với các dữ liệu lớn và đa dạng về tư thế, nhưng **không tốt** bằng CNN trong việc xử lý hình ảnh bị che khuất (occluded faces).
* **Khả năng mở rộng**: Transformer có khả năng **mở rộng tốt**, đặc biệt với dữ liệu lớn và phức tạp, nhờ vào cơ chế tự chú ý (self-attention).

### **Ưu điểm của Transformer trong FR:**

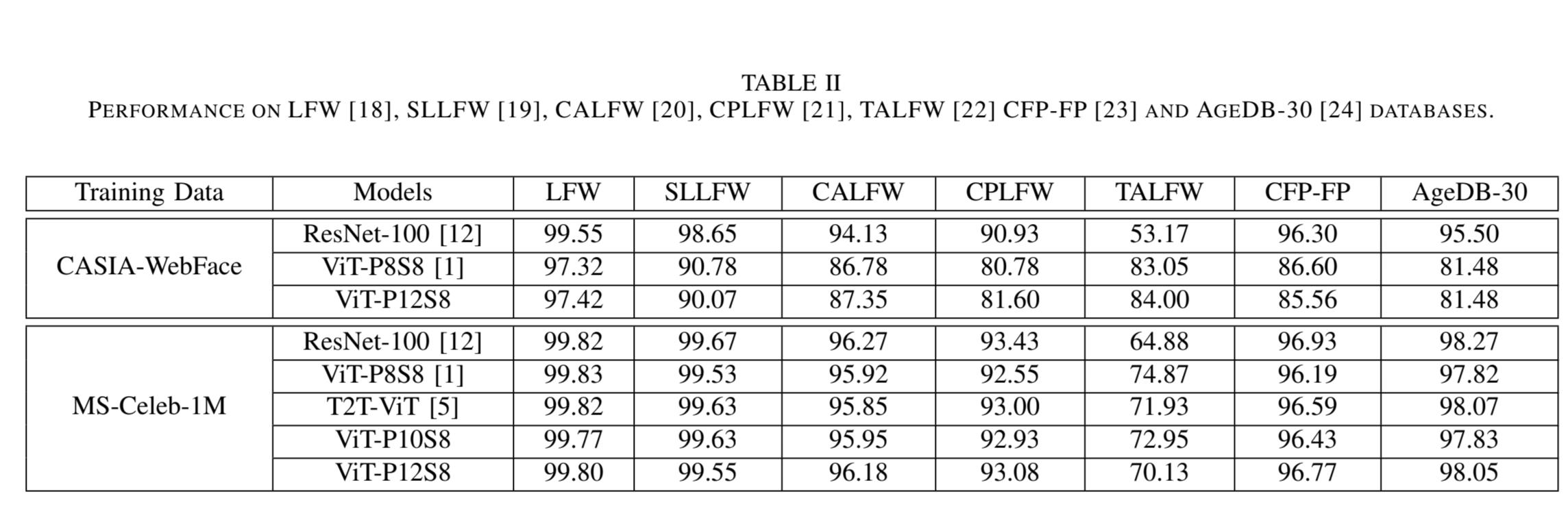
* **Mô hình hóa thông tin liên vùng**: Với cơ chế tự chú ý, Transformer có khả năng mô hình hóa mối quan hệ giữa các vùng trong ảnh khuôn mặt (các vùng khác nhau như mắt, mũi, miệng) tốt hơn so với CNN. Việc chồng lấn patch cải thiện thêm điều này.
* **Khả năng học tốt với dữ liệu lớn**: Khi huấn luyện với các tập dữ liệu lớn, mô hình Transformer có thể cạnh tranh về độ chính xác so với CNN với cùng số lượng tham số.
* **Phân tích vùng chú ý**: Transformer tự động tập trung vào các vùng quan trọng của khuôn mặt, giúp cải thiện khả năng nhận diện.

### **4. Nhược điểm của Transformer trong FR:**

* **Kém hiệu quả với dữ liệu nhỏ**: Khi sử dụng với tập dữ liệu nhỏ như CASIA-WebFace, Transformer gặp khó khăn trong việc tổng quát hóa và không đạt được kết quả tốt.
* **Tính phức tạp tính toán**: Cơ chế tự chú ý có độ phức tạp tính toán \(O(n^2)\), gây tốn kém tài nguyên khi độ dài chuỗi tăng, làm cho Transformer kém hiệu quả hơn khi phải xử lý một lượng lớn thông tin mà không có chiến lược giảm độ phức tạp.
* **Độ bền với che khuất**: Transformer không xử lý tốt các khuôn mặt bị che khuất (occluded faces) như CNN. Điều này có thể là do mô hình chú ý vào toàn bộ khuôn mặt, làm cho việc nhận diện trở nên khó khăn hơn khi chỉ một phần khuôn mặt bị hiển thị.



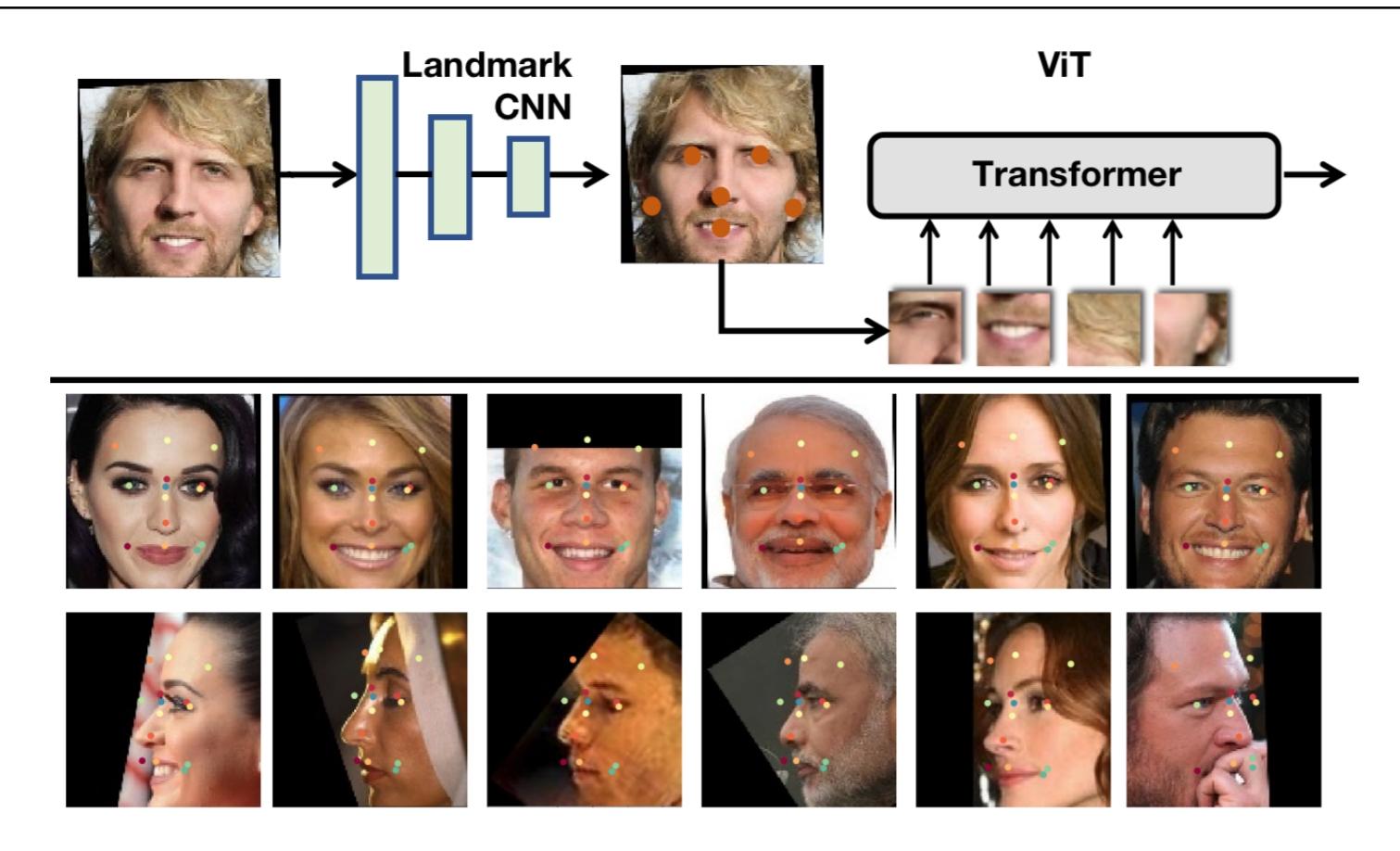
Bảng đánh giá hiệu xuất

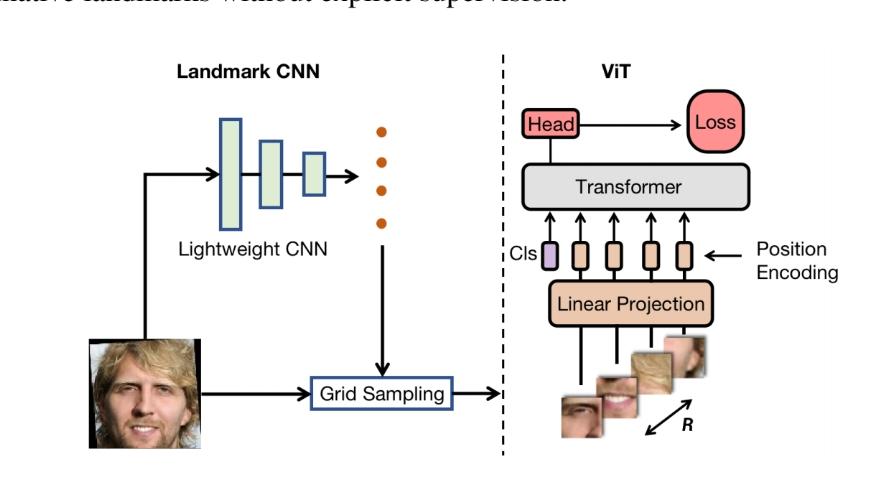


**Kết luận cuối cùng**

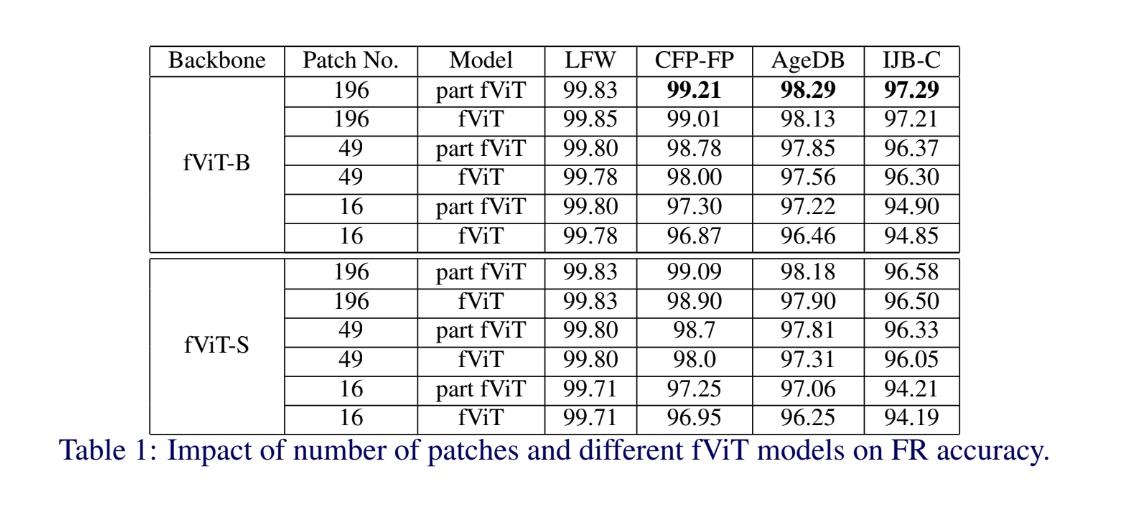
Việc áp dụng Transformer trong nhận diện khuôn mặt **có tính khả thi cao với dữ liệu lớn**, nhưng mô hình này **cần cải tiến để giải quyết** các **hạn chế về hiệu suất khi huấn luyện trên dữ liệu nhỏ** và **khả năng chống chịu với các yếu tố che khuất**. Transformer **có lợi thế trong** việc **mô hình hóa các mối quan hệ phức tạp giữa các phần của khuôn mặ**t, nhưng vẫn còn thách thức về tính hiệu quả tính toán và độ robust khi so sánh với CNNs trong một số trường hợp.

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





1. **Hiệu quả**:
   * **ViT** khi được huấn luyện đúng cách với các kỹ thuật tối ưu và dữ liệu tăng cường mạnh mẽ đã đạt được kết quả rất tốt, thể hiện qua các bộ dữ liệu phổ biến như **LFW**, **CFP-FP**, **IJB-B**, và **IJB-C**.
   * Phiên bản **part fViT**, sử dụng cách tiếp cận dựa trên phần khuôn mặt, giúp cải thiện hiệu suất hơn nữa, đặc biệt trên các bộ dữ liệu nhạy cảm với góc nhìn như **CFP-FP**. Điều này chứng tỏ khả năng của **Transformer** trong việc xử lý các đặc trưng không gian không đều và nhận diện các phần đặc trưng của khuôn mặt (như mắt, mũi, miệng).
2. **Tính khả thi**:
   * **Tính khả thi cao**: Việc huấn luyện **part fViT** có thể được thực hiện end-to-end, và mô hình không yêu cầu giám sát trực tiếp các **landmarks** trên khuôn mặt. Điều này làm cho mô hình dễ dàng triển khai trong các hệ thống nhận diện khuôn mặt thực tế mà không cần thông tin đầu vào phức tạp.
   * Các thử nghiệm với số lượng patches khác nhau cũng cho thấy rằng việc giảm số lượng **landmarks** giúp tăng tốc độ huấn luyện đáng kể mà không làm giảm quá nhiều hiệu suất. Ví dụ, mô hình với 49 **landmarks** nhanh hơn 4 lần so với mô hình với 196 **landmarks**, nhưng vẫn duy trì hiệu suất khá cao.
3. **Ưu điểm**:
   * **Khả năng nhận diện chính xác cao**: Cả **ViT** và **part fViT** đều đạt kết quả tiên tiến trên các benchmark quan trọng, với độ chính xác xác minh (verification accuracy) lên đến 99.83%.
   * **Xử lý phần khuôn mặt**: **part fViT** tận dụng cơ chế **attention** của **Transformer** để tập trung vào các phần quan trọng của khuôn mặt (như mắt, mũi, miệng), giúp mô hình học được các đặc trưng phân biệt mạnh mẽ hơn.
   * **Khả năng sử dụng trên lưới không đều**: **ViT** có thể xử lý các **visual tokens** từ các phần không đều, điều này khác với **CNNs** truyền thống chỉ hoạt động trên các lưới hình ảnh đều.
4. **Nhược điểm**:
   * **Overfitting**: Mô hình **ViT và part fViT**  dễ bị overfitting nếu không sử dụng các kỹ thuật tăng cường dữ liệu (data augmentation) phù hợp. Trong bài báo, nhóm tác giả đã phải áp dụng các kỹ thuật như **stochastic depth regularization (bỏ qua ngẫu nhiên một số layer)**, **RandAugment (Chọn ngẫu nhiên một thao tác xoay, lật …)**, **Cutout (cắt đi ngẫu nhiên một phần nào đó của ảnh đầu vào)** , và **Mixup (trộn hai ảnh)** để khắc phục điều này.
   * **Yêu cầu về tài nguyên**: **Transformer** thường yêu cầu tài nguyên tính toán lớn hơn so với các phương pháp **CNN** truyền thống, đặc biệt khi tăng số lượng **patches** hoặc **landmarks**. Điều này có thể khiến quá trình huấn luyện tốn kém hơn về mặt thời gian và tài nguyên.
   * **Khả năng bị suy giảm trên các bộ dữ liệu nhỏ**: Khi huấn luyện trên các bộ dữ liệu nhỏ hơn, như **CASIA-Webface**, **ViT** và **part fViT** cho kết quả kém hơn so với các mô hình **ResNet64** sử dụng các hàm mất mát tiên tiến hơn như **ArcFace**.



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

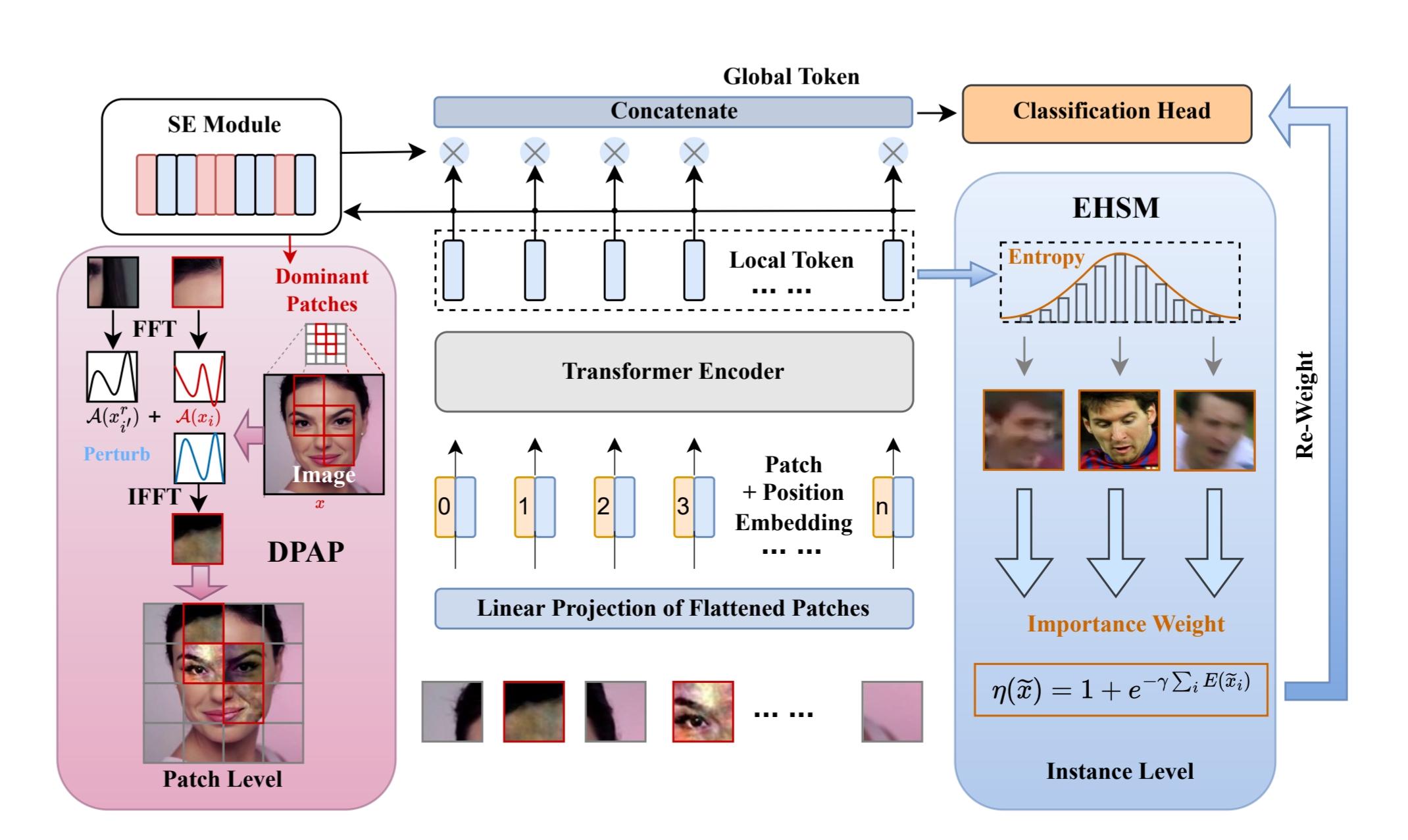
**Những vấn đề chính:**

* **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**

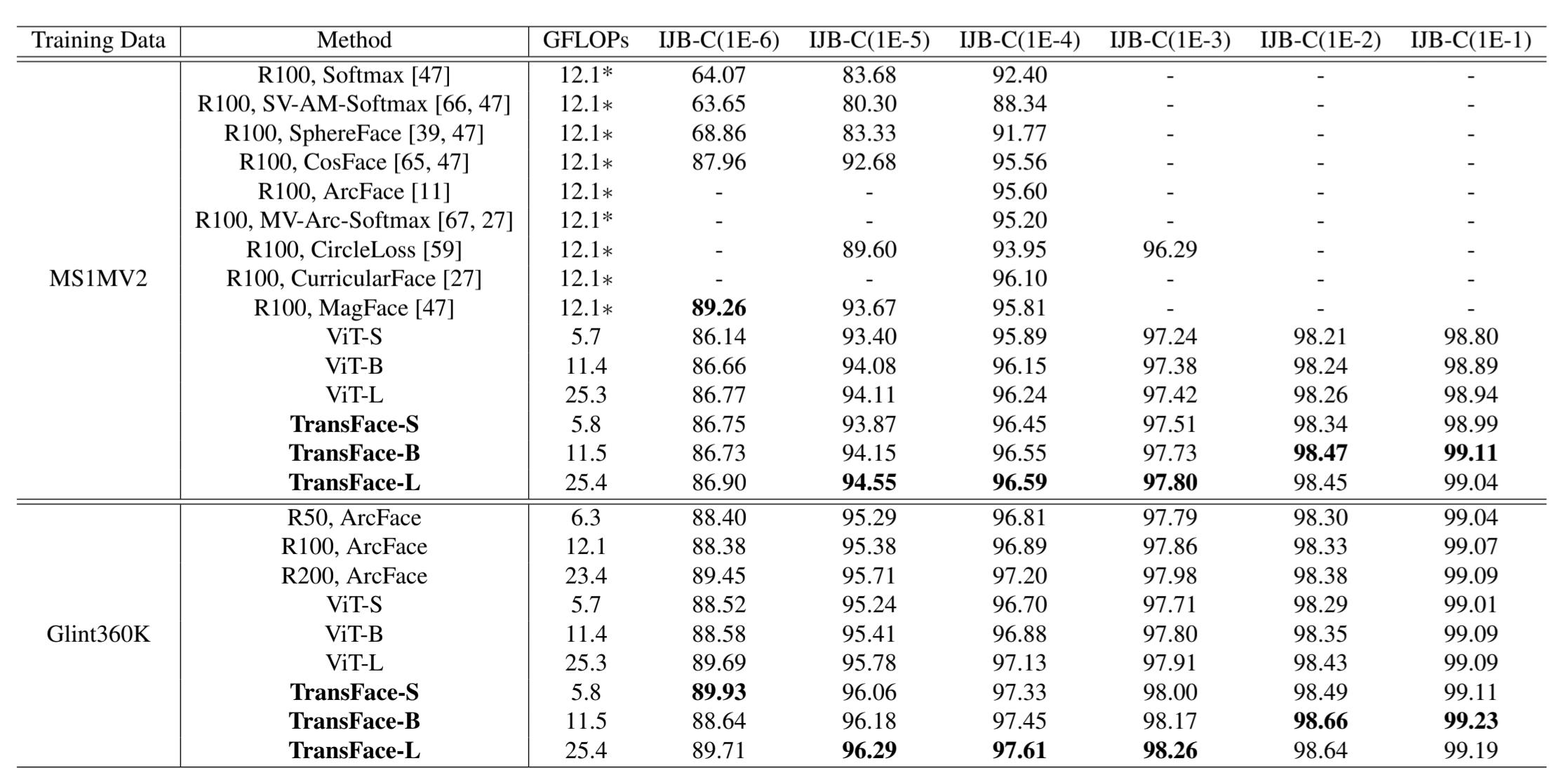
Công nghệ **hard sample mining** được sử dụng để tăng cường hiệu suất của mô hình bằng **cách khuyến khích mô hình học từ các mẫu khó hơn.** Các phương pháp khai thác mẫu khó trước đây chủ yếu sử dụng các chỉ số ở cấp độ mẫu, chẳng hạn như xác suất dự đoán hoặc **prediction loss**. Tuy nhiên, trong ViTs, 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.

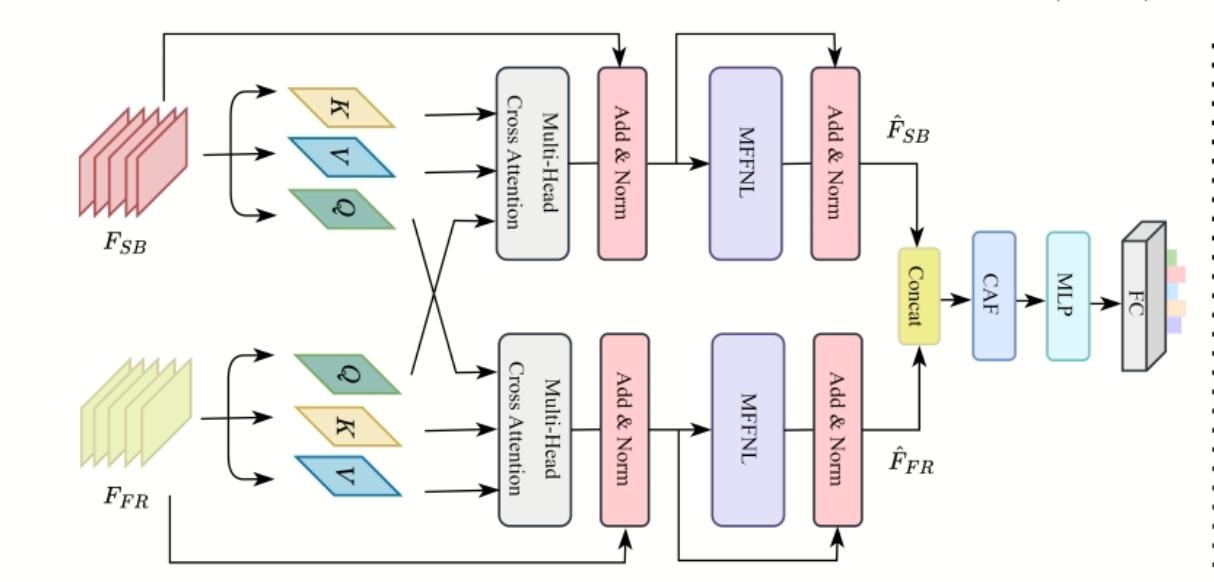
### **Phân tích và Nghiên cứu**

1. Kết quả cho thấy **ViT-S + DPAP** vượt trội hơn so với **ViT-S + SE**, cho thấy việc làm nhiễu biên độ các **dominant patches** có thể giảm bớt hiện tượng **overfitting**. **TransFace-S** cũng hoạt động tốt hơn **ViT-S + DPAP**, chứng tỏ chiến lược **EHSM** hiệu quả.
2. **So sánh với các chiến lược tăng cường dữ liệu trước đây:** Nhóm tác giả so sánh **DPAP** với các chiến lược tăng cường dữ liệu khác như **Random Erasing**, **Mixup**, **CutMix**, và **PatchErasing**. Kết quả cho thấy **DPAP** mang lại hiệu suất cao hơn cho **ViT**, nhờ vào việc sử dụng thông tin cấu trúc khuôn mặt và vị trí của các **dominant patches**.
3. **Khả năng cải thiện của EHSM:** Nhóm tác giả nghiên cứu xu hướng **entropy** của các **local tokens** trong quá trình huấn luyện. Kết quả cho thấy khi thêm chiến lược **EHSM**, lượng thông tin của các **token** tăng lên, chứng tỏ **EHSM** cải thiện khả năng biểu diễn đặc trưng của từng **local token**.
4. **Hiệu quả của EHSM:** Nhóm tác giả so sánh **EHSM** với các chiến lược **hard sample mining** khác như **ATk loss**, **MV-Softmax**, và **Focal loss**. **EHSM** vượt trội hơn các chiến lược trước đây, chứng minh rằng **EHSM** có thể đo lường độ khó của mẫu tốt hơn và nâng cao hiệu suất của mô hình.
5. **Trực quan hóa DPAP:** Nhóm tác giả trực quan hóa các mẫu huấn luyện ban đầu và các mẫu được tăng cường bởi **DPAP**. Các **dominant patches** chủ yếu phân bố gần tóc, trán và mắt, nhưng **DPAP** giúp mô hình sử dụng thêm các thông tin khác từ mũi, miệng, tai, và cằm, tăng cường khả năng tổng quát của mô hình.

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

Việc nhận diện chính xác khuôn mặt trong các hình ảnh chất lượng thấp vẫn còn là một thách thức do các chi tiết khuôn mặt bị che khuất. Tuy nhiên, có thể dự đoán một số thuộc tính sinh trắc học mềm (SB) như giới tính và độ hói ngay cả trong các hình ảnh chất lượng thấp. Nghiên cứu này đề xuất một mạng nơ-ron đa nhánh mới, sử dụng thông tin thuộc tính SB để nâng cao hiệu suất FR. Một mô-đun hợp nhất transformer hướng dẫn theo thuộc tính chéo (CATF) được giới thiệu nhằm nắm bắt các phụ thuộc và mối quan hệ dài hạn giữa các đại diện đặc trưng FR và SB

****

****

Để thực hiện điều này, nhóm nghiên cứu giới thiệu một mô-đun cross-attribute-guided transformer fusion (CATF), giúp bắt giữ các phụ thuộc dài hạn và mối quan hệ giữa các biểu diễn đặc trưng của FR và SB. Sự kết hợp thông tin qua quá trình dual cross-attention của mô-đun CATF giúp cải thiện hiệu suất FR. Ngoài ra, nhóm còn giới thiệu một khung self-attention distillation mới, giúp mạng neural chú ý đến các vùng quan trọng trên khuôn mặt, chẳng hạn như các đặc điểm nổi bật (landmarks), bằng cách căn chỉnh hình ảnh chất lượng thấp với hình ảnh chất lượng cao trong không gian đặc trưng. Quá trình này giúp mạng học được một biểu diễn đặc trưng không thay đổi theo chất lượng hình ảnh trong các môi trường không bị giới hạn (unconstrained environments).

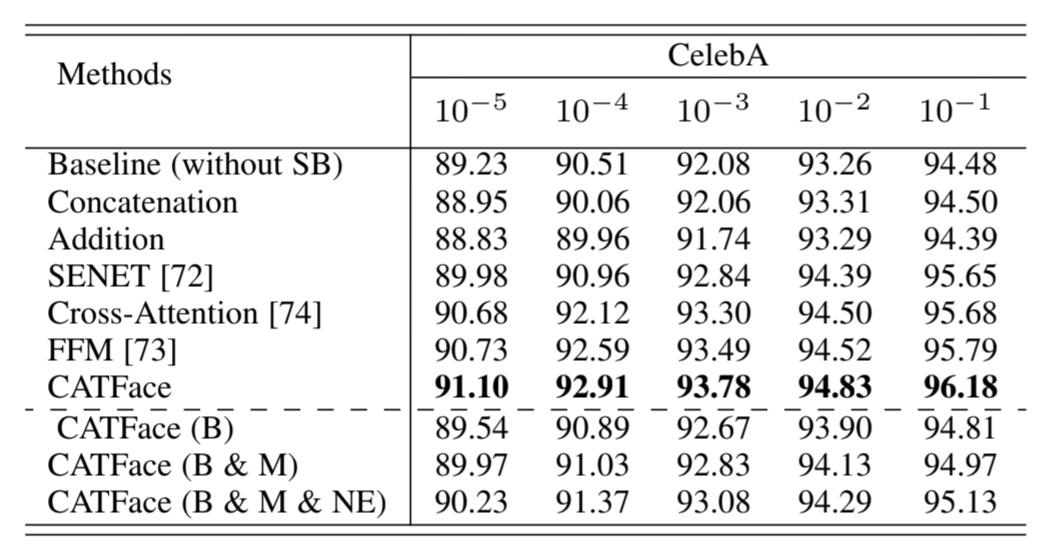
Nhóm đã tiến hành nhiều thí nghiệm trên các bộ dữ liệu FR khác nhau với chất lượng đa dạng, và kết quả thực nghiệm cho thấy phương pháp FR này vượt trội so với các nghiên cứu tiên tiến trong lĩnh vực này.

## Ưu điểm

* **Tăng cường hiệu suất nhận diện khuôn mặt**: Mô hình CATFace sử dụng thông tin thuộc tính sinh trắc học mềm (SB) như giới tính và độ hói để cải thiện khả năng nhận diện khuôn mặt trong các hình ảnh chất lượng thấp, giúp tăng độ chính xác trong các điều kiện khó khăn.
* **Mô-đun CATF**: Mô-đun hợp nhất transformer hướng dẫn theo thuộc tính chéo (CATF) hiệu quả trong việc nắm bắt các phụ thuộc và mối quan hệ dài hạn giữa các đặc trưng của FR và SB, tạo ra sự tương tác mạnh mẽ giữa các đại diện đặc trưng.
* **Khung chưng cất tự chú ý**: Mô hình áp dụng phương pháp chưng cất tự chú ý để làm nổi bật các vùng khuôn mặt quan trọng, giúp mạng học được các đặc trưng không phụ thuộc vào chất lượng hình ảnh.
* **Khả năng tổng quát tốt hơn**: Thực hiện nhiều thí nghiệm trên các bộ dữ liệu khác nhau cho thấy mô hình có khả năng tổng quát tốt hơn so với các phương pháp hiện tại, nhờ vào việc sử dụng dữ liệu huấn luyện đa dạng với nhiều biến thể thực tế.

## Nhược điểm

* **Chi phí tính toán cao**: Việc áp dụng các phương pháp tiền xử lý và mô-đun phức tạp có thể dẫn đến chi phí tính toán cao hơn so với một số phương pháp nhận diện khuôn mặt truyền thống.
* **Phụ thuộc vào chất lượng dữ liệu huấn luyện**: Mặc dù mô hình có khả năng xử lý hình ảnh chất lượng thấp, nhưng hiệu suất vẫn phụ thuộc vào chất lượng của dữ liệu huấn luyện và sự đa dạng của các thuộc tính SB được sử dụng.
* **Khó khăn trong việc xử lý các biến thể cực đoan**: Mặc dù mô hình đã cải thiện đáng kể khả năng nhận diện trong điều kiện khó khăn, nhưng việc duy trì độ chính xác cao khi gặp phải các biến thể cực đoan như góc nhìn rất nghiêng hoặc ánh sáng cực kỳ yếu vẫn là một thách thức lớn.



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

**Ưu điểm của KP-PRE:**

**Xử lý tốt các biến đổi về tỷ lệ và tư thế**:

KP-PRE tính toán độ lệch chú ý dựa trên khoảng cách tới các điểm mốc (keypoints) trên khuôn mặt, giúp mô hình thích nghi tốt hơn với các thay đổi về tỷ lệ (khuôn mặt to/nhỏ) và tư thế (góc xoay của khuôn mặt). Điều này cải thiện độ chính xác của nhận diện khuôn mặt trong các tình huống thực tế khi khuôn mặt có thể xoay, nghiêng hoặc ở khoảng cách khác nhau.

**Tăng độ chính xác của mô hình attention-based:**

Việc mã hóa vị trí tương đối giữa các keypoints giúp mô hình chú ý tốt hơn vào các vùng quan trọng của khuôn mặt, tăng khả năng phân biệt giữa các khuôn mặt tương tự nhau. Mô hình không chỉ dựa vào sự xuất hiện của các đặc trưng mà còn quan tâm đến mối quan hệ không gian giữa các điểm này, giúp nâng cao độ chính xác.

**Khả năng tích hợp vào các mô hình hiện tại**:

KP-PRE có thể dễ dàng tích hợp vào các kiến trúc Transformer hiện có mà không cần thay đổi nhiều về cấu trúc mô hình, chỉ điều chỉnh phần mã hóa vị trí. Điều này giúp các mô hình có thể tận dụng lợi thế của KP-PRE mà không đòi hỏi thiết kế lại toàn bộ hệ thống.

**Giảm nhiễu trong quá trình mã hóa vị trí:**

Bằng cách sử dụng khoảng cách đến các keypoints làm thông tin bổ trợ khi mã hóa vị trí, KP-PRE giảm thiểu ảnh hưởng của các đặc trưng nhiễu hoặc các thay đổi không quan trọng trong ảnh. Điều này giúp mô hình chú trọng vào các thông tin quan trọng để nhận diện.

**Nhược điểm của KP-PRE:**

Phụ thuộc vào việc xác định keypoints:

Hiệu quả của KP-PRE phụ thuộc rất nhiều vào việc xác định chính xác các điểm mốc (keypoints) trên khuôn mặt. Nếu các điểm này không được phát hiện chính xác (ví dụ như trong các hình ảnh mờ hoặc thiếu dữ liệu), hiệu suất của KP-PRE có thể bị suy giảm.

**Chi phí tính toán cao hơn**:

Việc thêm thông tin về khoảng cách từ các keypoints có thể làm tăng chi phí tính toán so với các phương pháp mã hóa vị trí truyền thống như RPE. Điều này có thể làm tăng độ phức tạp của mô hình và thời gian suy luận (inference time), đặc biệt trong các ứng dụng thời gian thực.

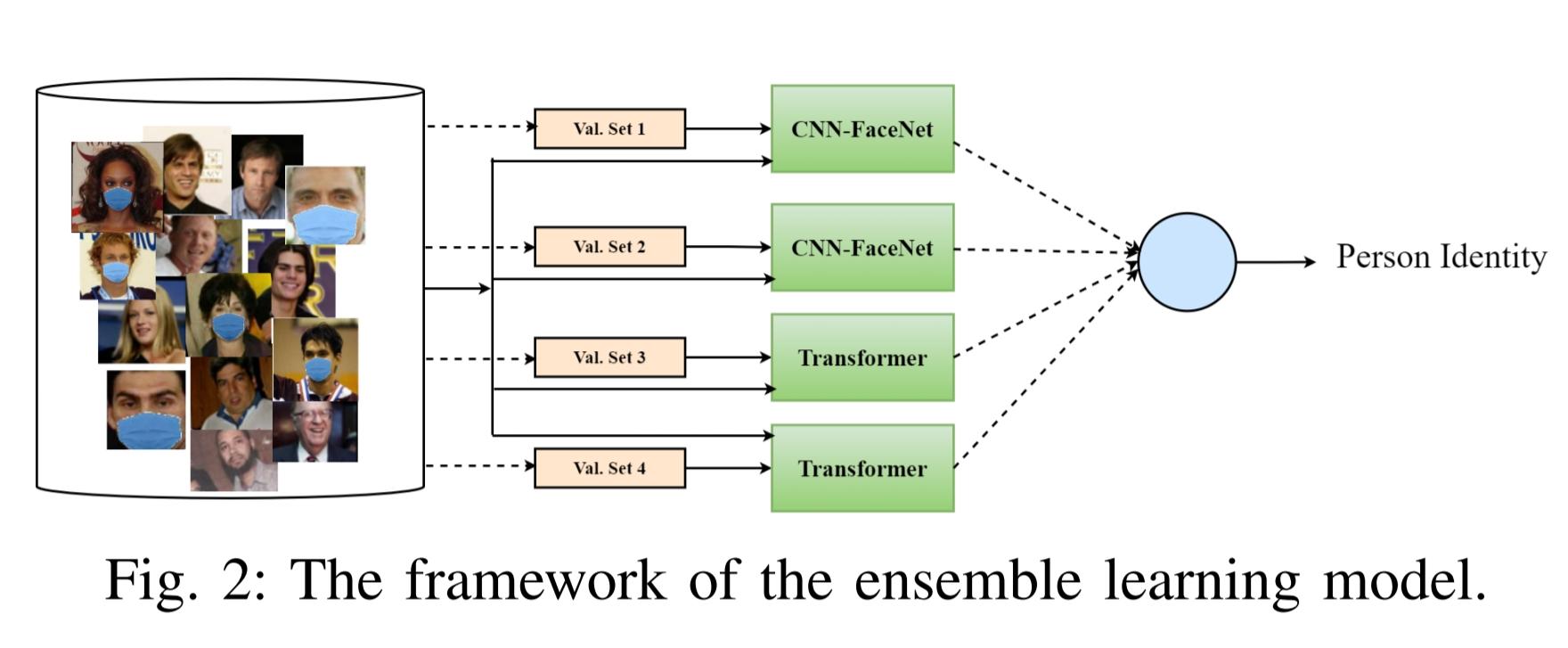
**Không hiệu quả trong trường hợp không có đủ thông tin keypoints**:

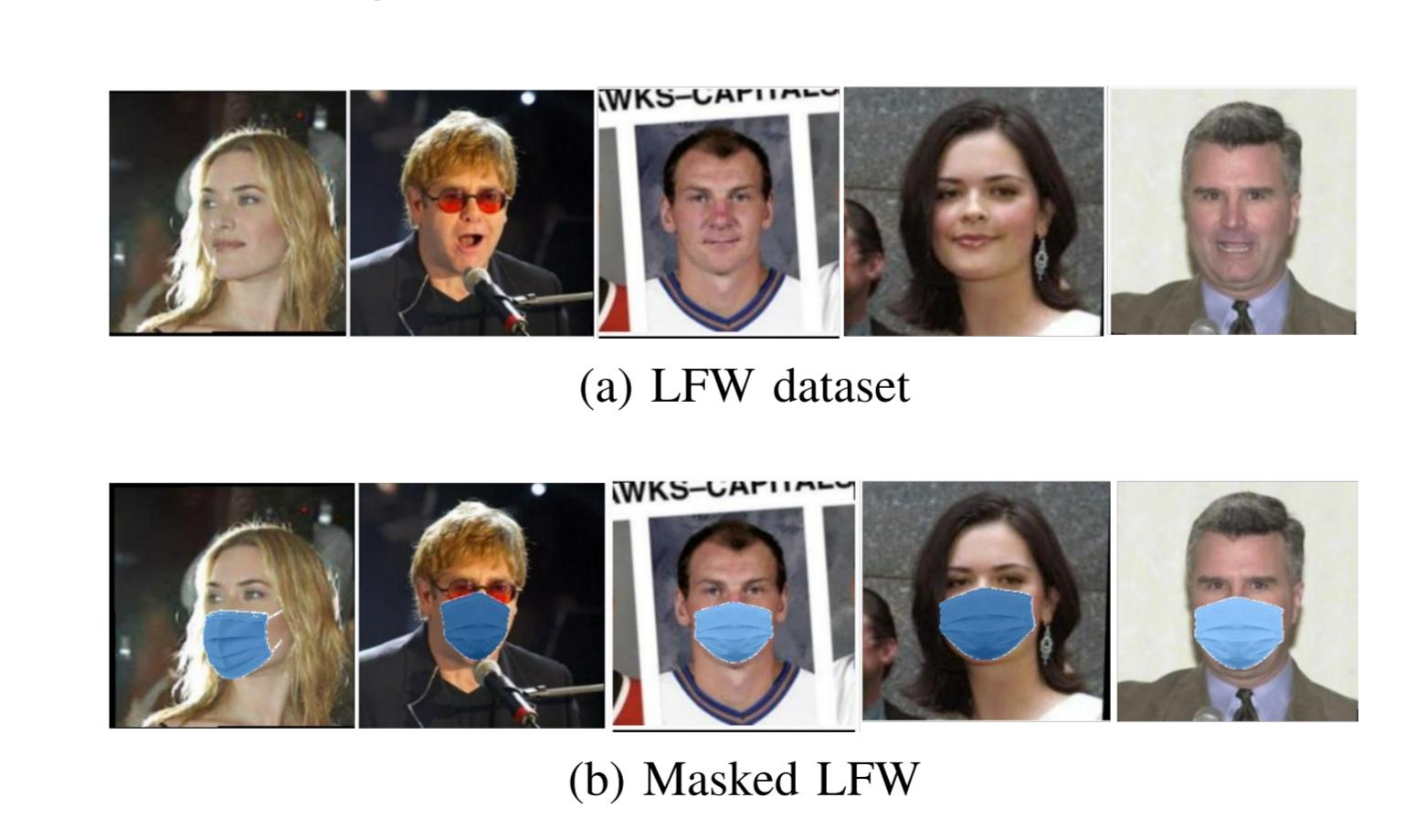
Nếu hình ảnh khuôn mặt không có đủ thông tin hoặc keypoints bị che khuất (ví dụ khi khuôn mặt bị che một phần hoặc chụp dưới điều kiện ánh sáng kém), KP-PRE có thể không hoạt động tốt, do nó phụ thuộc vào các điểm mốc này để tính toán vị trí tương đối.

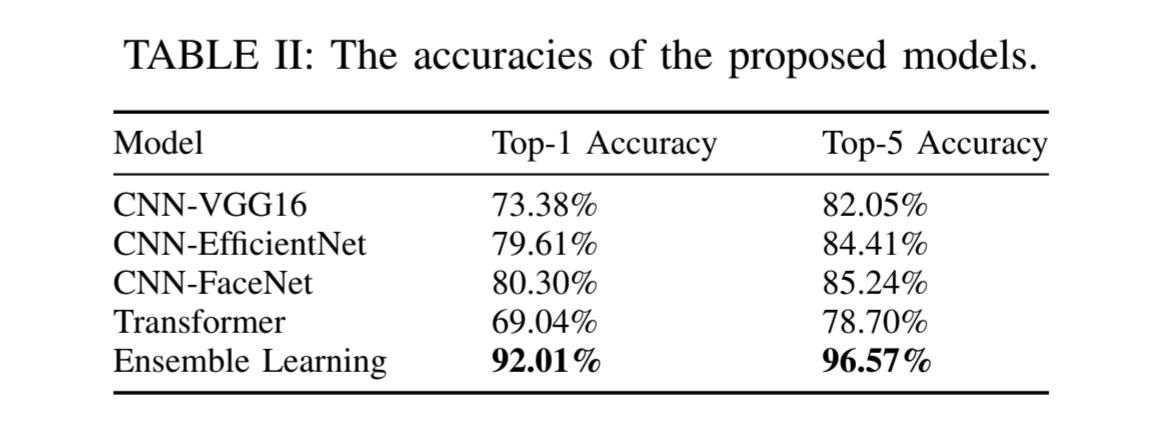
**Phức tạp trong triển khai**:

Mặc dù KP-PRE có thể tích hợp vào các mô hình Transformer hiện có, nhưng việc triển khai nó đòi hỏi phải có kiến thức sâu về cách thức hoạt động của attention và các kỹ thuật mã hóa vị trí. Điều này có thể là một rào cản đối với những nhóm nghiên cứu hoặc ứng dụng không quen thuộc với kỹ thuật này.

**Bài 6: Nhận diện khuôn mặt với khẩu trang**

Sử dụng **2 mô hình CNN** và **2 mô hình transformer** sau đó sử dụng kỹ thuật **bỏ phiếu đa số** để đưa ra dự đoán cuối cùng.





Mô-đun MaskTheFace đã được sử dụng trong công trình này để tạo phiên bản LFW với những gương mặt bị che khẩu trang. Mô-đun này tận dụng bộ phát hiện điểm mốc gương mặt để phát hiện sáu điểm chính trên khuôn mặt cần thiết cho việc áp dụng khẩu trang.

Kết quả thu được từ các thí nghiệm cho thấy rằng hệ thống đề xuất có thể đạt được độ chính xác cao trong việc nhận diện gương mặt bị che so với các phương pháp khác. Đặc biệt, mô hình kết hợp cho thấy hiệu suất vượt trội hơn so với từng mô hình riêng lẻ, nhờ vào khả năng tận dụng sức mạnh của nhiều mô hình khác nhau thông qua kỹ thuật bỏ phiếu đa số.

## Tóm tắt nội dung chính:

1. **Giới thiệu**:
   * Việc đeo khẩu trang đã trở thành một phần thiết yếu trong cuộc sống hàng ngày để ngăn chặn sự lây lan của coronavirus.
   * Các hệ thống nhận diện khuôn mặt hiện tại có độ chính xác cao với các khuôn mặt không bị che khuất nhưng gặp khó khăn khi nhận diện khuôn mặt có đeo khẩu trang.
2. **Hệ thống đề xuất**:
   * Hệ thống bao gồm hai mô hình mạng nơ-ron tích chập (CNN) và hai mô hình Transformer.
   * Các mô hình CNN được tinh chỉnh từ mô hình FaceNet đã được huấn luyện trước.
   * Dự đoán từ bốn mô hình được kết hợp bằng phương pháp bỏ phiếu đa số để xác định danh tính người đeo khẩu trang.
3. **Kết quả**:
   * Hệ thống được đánh giá trên tập dữ liệu LFW (Labeled Faces in the Wild) đã được tạo ra phiên bản có đeo khẩu trang.
   * Độ chính xác tốt nhất đạt được là 92%, vượt trội hơn so với các mô hình khác.
4. **Phân tích các phương pháp liên quan**:
   * Các phương pháp nhận diện khuôn mặt có đeo khẩu trang được chia thành hai loại: không phục hồi khuôn mặt (de-occlusion) và phục hồi trước khi nhận diện.
   * Nhiều nghiên cứu đã chỉ ra rằng việc phục hồi các vùng bị che khuất có thể cải thiện độ chính xác của nhận diện.
5. **Mô hình đề xuất**:
   * Ba mô hình CNN được tinh chỉnh từ VGG16, EfficientNet, và FaceNet.
   * Mô hình Transformer chỉ sử dụng thành phần mã hóa (encoder) để xử lý nhận diện khuôn mặt.
   * Sử dụng học tập tập hợp (ensemble learning) để kết hợp dự đoán từ các mô hình khác nhau nhằm cải thiện độ chính xác.
6. **Thí nghiệm**:
   * Tập dữ liệu LFW gốc không chứa khuôn mặt có đeo khẩu trang, do đó, một phiên bản mới với khẩu trang đã được tạo ra bằng cách sử dụng mô hình MaskTheFace để áp dụng khẩu trang lên các khuôn mặt trong LFW.

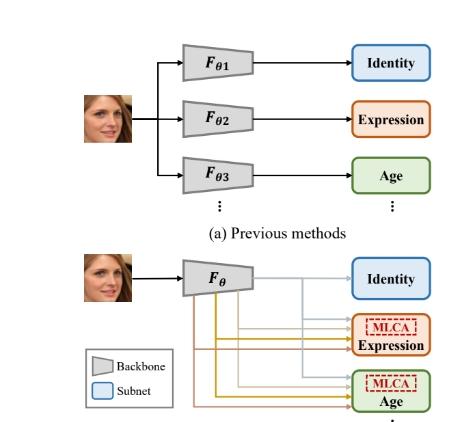
## Ưu điểm:

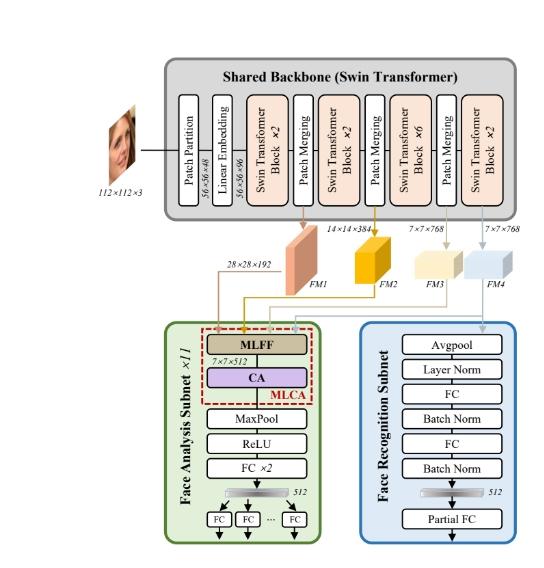
1. **Độ chính xác cao**: Mô hình đạt được độ chính xác lên tới 92% trong việc nhận diện khuôn mặt có đeo khẩu trang, vượt trội hơn so với nhiều phương pháp hiện tại.
2. **Sử dụng phương pháp học tập tập hợp (ensemble learning)**: Kết hợp dự đoán từ hai mô hình CNN và hai mô hình Transformer giúp cải thiện độ chính xác và độ tin cậy của hệ thống.
3. **Tối ưu hóa từ các mô hình đã được huấn luyện trước**: Mô hình CNN được tinh chỉnh từ các mô hình như FaceNet, VGG16 và EfficientNet, cho phép tận dụng các đặc trưng mạnh mẽ mà không cần huấn luyện từ đầu.
4. **Khả năng xử lý khuôn mặt bị che khuất**: Mô hình có khả năng nhận diện khuôn mặt ngay cả khi có sự che khuất lớn do khẩu trang, điều này rất quan trọng trong bối cảnh đại dịch COVID-19.
5. **Tính khả thi trong thực tế**: Hệ thống có thể triển khai trong các ứng dụng an ninh và nhận diện người dùng mà không cần yêu cầu người dùng bỏ khẩu trang.

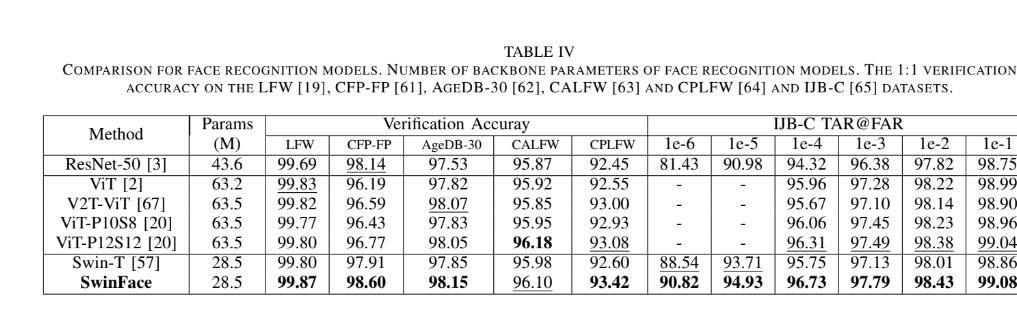
## Nhược điểm:

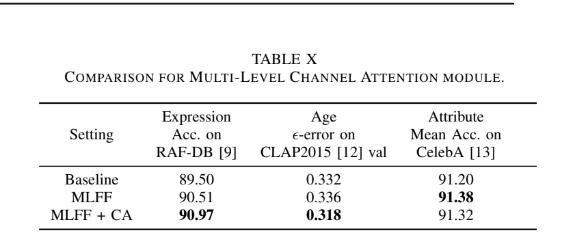
1. **Phụ thuộc vào chất lượng dữ liệu**: Độ chính xác của mô hình có thể giảm nếu dữ liệu đầu vào không đủ đa dạng hoặc chất lượng kém, đặc biệt là trong các tình huống ánh sáng khác nhau hoặc góc nhìn khác nhau.
2. **Tính phức tạp của mô hình**: Việc kết hợp nhiều mô hình có thể làm tăng độ phức tạp và thời gian tính toán, dẫn đến yêu cầu về tài nguyên phần cứng cao hơn.
3. **Khó khăn trong việc xử lý các biến thể khác nhau của khẩu trang**: Mô hình có thể gặp khó khăn khi đối mặt với các kiểu khẩu trang khác nhau hoặc khi chúng được đeo không đúng cách.
4. **Chưa hoàn toàn giải quyết vấn đề nhận diện cảm xúc**: Việc che khuất một phần lớn khuôn mặt có thể ảnh hưởng đến khả năng nhận diện cảm xúc, điều này có thể hạn chế ứng dụng của mô hình trong các lĩnh vực yêu cầu phân tích cảm xúc.
5. **Khả năng mở rộng**: Mô hình hiện tại chủ yếu tập trung vào việc nhận diện khuôn mặt có đeo khẩu trang; việc mở rộng sang các nhiệm vụ khác như nhận diện biểu cảm hoặc thuộc tính khuôn mặt vẫn cần nghiên cứu thêm.

**Bài 7: Swin Transformer**

****

****

****

****

SwinFace: Một Transformer Đa Nhiệm cho Nhận Diện Khuôn Mặt, Nhận Diện Biểu Cảm, Ước Lượng Tuổi và Ước Lượng Thuộc TínhLixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, Weihong Deng, Thành viên IEEETóm tắt—Trong những năm gần đây, các transformer thị giác đã được giới thiệu vào nhận diện và phân tích khuôn mặt và đã đạt được những bước đột phá về hiệu suất. Tuy nhiên, hầu hết các phương pháp trước đây thường huấn luyện một mô hình đơn hoặc một tập hợp các mô hình để thực hiện nhiệm vụ mong muốn, điều này bỏ qua sự tương tác giữa các nhiệm vụ khác nhau và không đạt được độ chính xác dự đoán cải thiện, tăng hiệu quả dữ liệu và giảm thời gian huấn luyện. Bài báo này trình bày một thuật toán đa mục đích cho việc nhận diện khuôn mặt đồng thời, nhận diện biểu cảm khuôn mặt, ước lượng tuổi và ước lượng thuộc tính khuôn mặt (40 thuộc tính bao gồm giới tính) dựa trên một Swin Transformer duy nhất. Thiết kế của chúng tôi, SwinFace, bao gồm một backbone chung duy nhất cùng với một subnet cho mỗi tập hợp nhiệm vụ liên quan. Để giải quyết các xung đột giữa nhiều nhiệm vụ và đáp ứng các yêu cầu khác nhau của nhiệm vụ, một mô-đun Chú Ý Kênh Đa Cấp (MLCA) được tích hợp vào mỗi subnet phân tích cụ thể nhiệm vụ, có khả năng chọn lựa các đặc trưng từ các cấp độ và kênh tối ưu để thực hiện các nhiệm vụ mong muốn. Các thí nghiệm rộng rãi cho thấy mô hình được đề xuất có khả năng hiểu biết tốt hơn về khuôn mặt và đạt được hiệu suất xuất sắc cho tất cả các nhiệm vụ. **Đặc biệt, nó đạt được độ chính xác 90.97% trên RAF-DB và lỗi 0.22 ϵ trên CLAP2015, đây là những kết quả hàng đầu trong nhận diện biểu cảm khuôn mặt và ước lượng tuổi tương ứng.**

Từ khóa—Học đa nhiệm, Swin Transformer, Nhận diện khuôn mặt, Nhận diện biểu cảm khuôn mặt, Ước lượng tuổi, Ước lượng thuộc tính khuôn mặt.

I. GIỚI THIỆU

Nhận diện và phân tích khuôn mặt là những chủ đề quan trọng trong lĩnh vực thị giác máy tính và có một loạt ứng dụng trong giám sát an ninh, giải trí kỹ thuật số, nhận diện cảm xúc, v.v. Gần đây, các nhà nghiên cứu đã giới thiệu các transformer thị giác vào nhận diện và phân tích khuôn mặt và đã đạt được những bước đột phá về hiệu suất trên một số nhiệm vụ. Ví dụ, An et al.

đã chứng minh rằng các mạng dựa trên ViT có thể đạt được hiệu suất tốt hơn so với các mạng dựa trên ResNet trong nhận diện khuôn mặt; TransFER đã khám phá các transformer cho nhận diện biểu cảm khuôn mặt và đạt được cải thiện hiệu suất đáng kể. Tuy nhiên, các mô hình dựa trên transformer này thường được thiết kế để chỉ đạt được một nhiệm vụ cụ thể nào đó, điều này dẫn đến những hạn chế sau:

Tính khan hiếm dữ liệu. So với nhận diện khuôn mặt, các nhiệm vụ phân tích khuôn mặt như nhận diện biểu cảm khuôn mặt, ước lượng tuổi và phân loại thuộc tính khuôn mặt vẫn còn thách thức do thiếu dữ liệu huấn luyện lớn có sẵn.

Hiệu quả mô hình. Việc học các mạng riêng biệt cho các nhiệm vụ khác nhau sẽ dẫn đến sự không hiệu quả về bộ nhớ và tốc độ suy diễn.

Mặc dù học đa nhiệm đã được đề xuất trong các mạng nơ-ron tích chập (CNNs) để giải quyết những vấn đề này... nhưng vẫn là một lĩnh vực nghiên cứu chưa được khai thác nhiều trong các transformer.Trong bài báo này, chúng tôi huấn luyện một transformer đồng thời trong một khung học đa nhiệm mà giải quyết đồng thời các nhiệm vụ của nhận diện khuôn mặt, nhận diện biểu cảm khuôn mặt, ước lượng tuổi và ước lượng thuộc tính khuôn mặt (40 thuộc tính bao gồm giới tính). Thiết kế của chúng tôi, SwinFace, bao gồm một backbone Swin Transformer chung duy nhất cùng với một subnet nhận diện khuôn mặt và 11 subnet phân tích khuôn mặt.Như thể hiện trong Bảng II... để giảm tính toán, chúng tôi nhóm một số nhiệm vụ phân tích khuôn mặt liên quan lại với nhau và xử lý chúng bằng một subnet thay vì thiết kế một subnet cho mỗi nhiệm vụ. Bằng cách chia sẻ đại diện và tận dụng sự tương tác giữa các nhiệm vụ liên quan... kiến thức chứa đựng trong một nhiệm vụ có thể được sử dụng bởi các nhiệm vụ khác...Hình 1 so sánh phương pháp trước đây với phương pháp của chúng tôi. Ngoài ra... học đa nhiệm vốn dĩ là một vấn đề đa mục tiêu và các nhiệm vụ khác nhau có thể xung đột với nhau. Ví dụ... nếu hai subnet cụ thể cho từng nhiệm vụ này cùng phân nhánh từ lớp trên cùng... rõ ràng rằng chúng sẽ hạn chế lẫn nhau trong việc cải thiện hiệu suất do mục tiêu xung đột.Để tránh xung đột này,... chúng tôi đề xuất một mô-đun Chú Ý Kênh Đa Cấp (MLCA) được tích hợp vào mỗi subnet phân tích khuôn mặt...Bài báo này đóng góp những điểm sau:

Đây là mô hình đa nhiệm đầu tiên giải quyết đồng thời một tập hợp đa dạng các nhiệm vụ phân tích (42 nhiệm vụ) và nhận diện bằng cách sử dụng một transformer duy nhất.

Chúng tôi đề xuất mô-đun Chú Ý Kênh Đa Cấp (MLCA) để xử lý xung đột trích xuất đặc trưng của backbone và lựa chọn đặc trưng của subnets.

Mô hình đề xuất đạt độ chính xác 90.97% trên RAF-DB…

II. CÁC CÔNG TRÌNH LIÊN QUAN

Trong phần này...

A. Nhận Diện Khuôn Mặt

Nhận Diện Khuôn Mặt (FR) là một nhiệm vụ quan trọng trong thị giác máy tính...

B. Nhận Diện Biểu Cảm Khuôn Mặt

Nhận Diện Biểu Cảm Khuôn Mặt (FER)...

C. Ước Lượng Tuổi

Ước Lượng Tuổi (AE)...

D. Ước Lượng Thuộc Tính Khuôn Mặt

Ước Lượng Thuộc Tính Khuôn Mặt (FAE)...

E. Học Đa Nhiệm

Học Đa Nhiệm (MTL)...

III. PHƯƠNG PHÁP TIẾN HÀNH

A. Kiến Trúc Mô Hình

Hình 2 trình bày kiến trúc tổng thể của mô hình SwinFace. Mô hình bao gồm một backbone chung là Swin Transformer và nhiều subnet cho các nhiệm vụ khác nhau. Phần backbone thực hiện việc trích xuất đặc trưng từ đầu vào hình ảnh, sau đó các đặc trưng này được chia sẻ giữa các subnet khác nhau để thực hiện các nhiệm vụ nhận diện khuôn mặt, nhận diện biểu cảm, ước lượng tuổi và ước lượng thuộc tính khuôn mặt.

1) Backbone

Backbone của chúng tôi là một Swin Transformer, được thiết kế để xử lý các đặc trưng hình ảnh ở nhiều cấp độ khác nhau. Nó bao gồm các lớp phân đoạn patch, nhúng tuyến tính và nhiều khối Swin Transformer với khả năng xử lý thông tin cục bộ và toàn cục.

2) Subnet Nhận Diện Khuôn Mặt

Subnet nhận diện khuôn mặt sử dụng các đặc trưng từ backbone để thực hiện nhiệm vụ nhận diện khuôn mặt. Nó bao gồm các lớp Fully Connected (FC) với các hàm kích hoạt ReLU và Batch Normalization để cải thiện độ chính xác của mô hình.

3) Subnet Phân Tích Khuôn Mặt

Mỗi subnet phân tích khuôn mặt được thiết kế cho từng nhiệm vụ cụ thể như nhận diện biểu cảm, ước lượng tuổi và ước lượng thuộc tính. Mỗi subnet này sử dụng mô-đun Chú Ý Kênh Đa Cấp (MLCA) để tối ưu hóa việc lựa chọn đặc trưng từ các cấp độ khác nhau trong backbone.

B. Mô-đun Chú Ý Kênh Đa Cấp (MLCA)

Mô-đun MLCA bao gồm hai thành phần chính: Multi-Level Feature Fusion (MLFF) và Channel Attention (CA).

MLFF: Kết hợp các đặc trưng ở nhiều cấp độ khác nhau một cách hiệu quả, cho phép SwinFace dựa vào cả thông tin cục bộ và toàn cục để thực hiện phân tích khuôn mặt.

CA: Gán trọng số cho các đặc trưng từ các cấp độ khác nhau một cách thích ứng, cho phép từng nhiệm vụ tự lựa chọn đặc trưng phù hợp với yêu cầu của mình. Điều này giúp giải quyết vấn đề xung đột giữa các nhiệm vụ.

C. Huấn Luyện Mô Hình

Mô hình được huấn luyện trong một khung học đa nhiệm, nơi mà tất cả các subnet đều được tối ưu hóa đồng thời. Điều này không chỉ giúp cải thiện hiệu suất của từng nhiệm vụ mà còn tận dụng sự tương tác giữa chúng để nâng cao khả năng hiểu biết về khuôn mặt của mô hình.

D. Thí Nghiệm

Chúng tôi đã tiến hành nhiều thí nghiệm trên các tập dữ liệu khác nhau để đánh giá hiệu suất của mô hình SwinFace. Các kết quả cho thấy rằng mô hình không chỉ đạt được độ chính xác cao trong nhận diện biểu cảm mà còn cải thiện đáng kể trong ước lượng tuổi và thuộc tính khuôn mặt so với các phương pháp hiện có.

IV. KẾT QUẢ VÀ THẢO LUẬN

Kết quả thí nghiệm cho thấy mô hình SwinFace đạt được những thành tích xuất sắc trên nhiều nhiệm vụ phân tích khuôn mặt:

Độ chính xác: Mô hình đạt được độ chính xác 90.97% trên tập dữ liệu RAF-DB cho nhiệm vụ nhận diện biểu cảm.

Ước lượng tuổi: Mô hình đạt lỗi 0.22 trên tập dữ liệu CLAP2015, cho thấy khả năng ước lượng tuổi tốt.

Các thí nghiệm cũng chỉ ra rằng việc sử dụng một backbone chung giúp giảm thiểu chi phí tính toán và tăng hiệu quả sử dụng dữ liệu, đồng thời cải thiện tốc độ huấn luyện so với việc huấn luyện từng mô hình riêng biệt cho mỗi nhiệm vụ.

V. KẾT LUẬN

Bài báo này đã giới thiệu mô hình SwinFace, một transformer đa nhiệm cho nhận diện khuôn mặt, nhận diện biểu cảm, ước lượng tuổi và ước lượng thuộc tính khuôn mặt. Với thiết kế sáng tạo kết hợp giữa backbone chung và các subnet chuyên biệt cùng với mô-đun MLCA, SwinFace đã chứng minh khả năng vượt trội trong việc xử lý đồng thời nhiều nhiệm vụ phân tích khuôn mặt. Các kết quả thí nghiệm cho thấy rằng mô hình không chỉ cải thiện hiệu suất mà còn tối ưu hóa quy trình huấn luyện, mở ra hướng đi mới cho nghiên cứu trong lĩnh vực này.

Dịch phần 5: Experiments

V. THÍ NGHIỆM

A. Thiết Lập Thí Nghiệm

Chúng tôi đã tiến hành một loạt thí nghiệm để đánh giá hiệu suất của mô hình SwinFace trên các nhiệm vụ nhận diện khuôn mặt, nhận diện biểu cảm, ước lượng tuổi và ước lượng thuộc tính khuôn mặt. Các thí nghiệm được thực hiện trên nhiều tập dữ liệu khác nhau, bao gồm RAF-DB cho nhận diện biểu cảm, CLAP2015 cho ước lượng tuổi, và CelebA cho ước lượng thuộc tính.

B. Tập Dữ Liệu

Các tập dữ liệu được sử dụng trong các thí nghiệm bao gồm:

RAF-DB: Tập dữ liệu gồm 15,339 hình ảnh cho nhiệm vụ nhận diện biểu cảm khuôn mặt.

CLAP2015: Tập dữ liệu với 4,699 hình ảnh cho nhiệm vụ ước lượng tuổi.

CelebA: Tập dữ liệu chứa khoảng 200K hình ảnh với 40 thuộc tính khuôn mặt.

C. Phương Pháp Huấn Luyện

Mô hình SwinFace được huấn luyện trong một khung học đa nhiệm. Chúng tôi sử dụng hàm mất mát kết hợp cho tất cả các nhiệm vụ, nhằm tối ưu hóa đồng thời các subnet cho từng nhiệm vụ cụ thể. Quá trình huấn luyện được thực hiện trên GPU với kích thước batch là 64 và sử dụng AdamW làm thuật toán tối ưu.

D. Kết Quả Thí Nghiệm

Kết quả thí nghiệm cho thấy mô hình SwinFace đạt được những thành tích xuất sắc trên tất cả các nhiệm vụ:

Nhận Diện Biểu Cảm: Mô hình đạt độ chính xác 90.97% trên tập dữ liệu RAF-DB, vượt qua nhiều phương pháp hiện có.

Ước Lượng Tuổi: Mô hình đạt lỗi 0.22 trên tập CLAP2015, cho thấy khả năng ước lượng tuổi tốt.

Ước Lượng Thuộc Tính: Mô hình cũng thể hiện hiệu suất cao trong việc phân loại các thuộc tính khuôn mặt như giới tính và các đặc điểm khác.

E. Phân Tích Kết Quả

Kết quả cho thấy rằng việc sử dụng một backbone chung và mô-đun Chú Ý Kênh Đa Cấp (MLCA) giúp cải thiện hiệu suất của mô hình trong việc xử lý đồng thời nhiều nhiệm vụ phân tích khuôn mặt. Sự tương tác giữa các nhiệm vụ cũng đóng góp vào việc cải thiện độ chính xác và giảm thời gian huấn luyện.

F. So Sánh Với Các Phương Pháp Khác

Chúng tôi so sánh kết quả của mô hình SwinFace với một số phương pháp hiện có trong bảng dưới đây:

Nhiệm VụMô Hình Hiện CóĐộ Chính Xác (%)Nhận Diện Biểu CảmTransFER89.50Ước Lượng TuổiBridgeNet0.25Ước Lượng Thuộc TínhMCFA92.00

Kết quả này chứng minh rằng SwinFace không chỉ vượt trội về hiệu suất mà còn mang lại lợi ích về hiệu quả tính toán khi xử lý đồng thời nhiều nhiệm vụ.

G. Kết Luận Thí Nghiệm

Các thí nghiệm đã chứng minh rằng mô hình SwinFace có khả năng hiểu biết tốt hơn về khuôn mặt và đạt được hiệu suất xuất sắc cho tất cả các nhiệm vụ phân tích khuôn mặt. Việc áp dụng học đa nhiệm không chỉ cải thiện độ chính xác mà còn tối ưu hóa quy trình huấn luyện, mở ra hướng đi mới cho nghiên cứu trong lĩnh vực này.

**Ưu điểm của mô hình SwinFace:**

* **Hiệu suất cao trên nhiều nhiệm vụ phân tích khuôn mặt** như nhận diện biểu cảm (90.97% trên RAF-DB), ước lượng tuổi (lỗi 0.22 trên CLAP2015), ước lượng thuộc tính khuôn mặt
* Tận dụng sự tương tác giữa các nhiệm vụ để cải thiện độ chính xác và **giảm thời gian huấn luyện**
* **Sử dụng backbone chung** giúp **giảm thiểu chi phí tính toán** và **tăng hiệu quả sử dụng dữ liệu**
* **Mô-đun Chú Ý Kênh Đa Cấp (MLCA)** cho phép các **subnet thích ứng lựa chọn đặc trưng phù hợp với từng nhiệm vụ**, giải quyết vấn đề xung đột giữa các nhiệm vụ

Mô-đun MLCA bao gồm hai thành phần chính: Multi-Level Feature Fusion (MLFF) và Channel Attention (CA).

MLFF: Kết hợp các đặc trưng ở nhiều cấp độ khác nhau một cách hiệu quả, cho phép SwinFace dựa vào cả thông tin cục bộ và toàn cục để thực hiện phân tích khuôn mặt.

CA: Gán trọng số cho các đặc trưng từ các cấp độ khác nhau một cách thích ứng, cho phép từng nhiệm vụ tự lựa chọn đặc trưng phù hợp với yêu cầu của mình. Điều này giúp giải quyết vấn đề xung đột giữa các nhiệm vụ.

* Mã nguồn và mô hình sẽ được công khai, mang lại lợi ích cho cộng đồng nghiên cứu

**Block swinTransformer gồm những gì**

Block của Swin Transformer bao gồm các thành phần chính sau:

Patch Partition: Chia hình ảnh đầu vào thành các patch nhỏ để xử lý.

Linear Embedding: Chuyển đổi các patch thành các vector nhúng.

Swin Transformer Blocks: Gồm nhiều khối transformer, mỗi khối thực hiện chú ý cục bộ và toàn cục. Các khối này được tổ chức theo các lớp khác nhau để trích xuất đặc trưng ở nhiều cấp độ.

Patch Merging: Kết hợp các patch lại với nhau để giảm kích thước đầu vào và tăng cường thông tin.

Layer Norm: Chuẩn hóa đầu ra của các lớp để cải thiện tính ổn định trong quá trình huấn luyện.

Fully Connected (FC) Layers: Các lớp kết nối đầy đủ dùng để phân loại và dự đoán thuộc tính khuôn mặt.

Multi-Level Channel Attention (MLCA): Một mô-đun chú ý cho phép lựa chọn đặc trưng từ các cấp độ khác nhau, tối ưu hóa việc trích xuất thông tin cho từng nhiệm vụ cụ thể