1. PART-BASED
2. Ngữ cảnh của bài toán là gì?

Bài báo chỉ ra rằng các phương pháp truyền thống trong nhận dạng khuôn mặt chủ yếu dựa vào mạng nơ-ron tích chập (CNN) để xử lý hình ảnh khuôn mặt một cách toàn diện. Tuy nhiên, tác giả cho rằng việc xử lý khuôn mặt như một cấu trúc gồm nhiều phần (như mắt, mũi, miệng) có thể mang lại hiệu quả cao hơn. Do đó, bài báo tập trung vào hai điểm chính:

* Sử dụng Vision Transformer: Tác giả áp dụng ViT như một kiến trúc mạnh mẽ cho việc nhận dạng khuôn mặt, cho thấy rằng nó có thể vượt qua nhiều phương pháp hiện tại trong lĩnh vực này.
* Phát triển quy trình part fViT: Quy trình này bao gồm việc sử dụng một mạng CNN nhẹ để dự đoán tọa độ của các điểm đặc trưng trên khuôn mặt, từ đó trích xuất các phần khuôn mặt và đưa vào ViT để nhận dạng. Điều đặc biệt là quy trình này không cần sự giám sát về tọa độ các điểm đặc trưng, điều này giúp đơn giản hóa quá trình huấn luyện.

1. Vấn đề gì đang gặp phải trong ngữ cảnh này của các pp cũ?
   * Phương pháp toàn diện (Holistic Methods): Các phương pháp truyền thống, chủ yếu dựa vào mạng nơ-ron tích chập (CNN), xử lý hình ảnh khuôn mặt một cách toàn diện. Điều này có nghĩa là chúng không tận dụng được cấu trúc phức tạp của khuôn mặt, dẫn đến việc bỏ qua thông tin quan trọng từ các phần riêng lẻ như mắt, mũi và miệng.
   * Thiếu khả năng phân tích các phần riêng lẻ: Các mô hình CNN thường sử dụng lưới không gian đồng nhất để trích xuất đặc trưng, điều này không phù hợp với cách mà khuôn mặt được cấu thành từ nhiều phần khác nhau. Việc này có thể làm giảm độ chính xác trong nhận dạng, nhất là khi khuôn mặt bị che khuất hoặc có sự biến đổi về góc nhìn.
   * Yêu cầu giám sát cho điểm đặc trưng: Nhiều phương pháp hiện tại yêu cầu phải có sự giám sát cho các điểm đặc trưng trên khuôn mặt, điều này làm cho quá trình huấn luyện trở nên phức tạp và tốn thời gian.
2. Họ đề xuất pp gì, giúp cải thiện vấn đề gì trong ngữ cảnh này? Chi tiết giải pháp của họ.
   * các module nào, đầu vào ra ntn, luồng chạy dữ liệu, thuật toán đề xuất chính là gì?
   * mindset tại sao họ lại làm ntn.

Cấu trúc của **part fViT**

* **Sử dụng Vision Transformer (ViT):** Phương pháp này tận dụng kiến trúc ViT để xử lý các phần khuôn mặt, cho phép mô hình hoạt động trên các mảnh hình ảnh không đều (irregular patches) thay vì lưới đều như trong các mạng nơ-ron tích chập (CNN) truyền thống.
* **Mạng CNN Nhẹ để Dự đoán Tọa độ Điểm Đặc Trưng**: Một mạng CNN nhẹ được sử dụng để dự đoán tọa độ của các điểm đặc trưng trên khuôn mặt. Các tọa độ này sau đó được sử dụng để trích xuất các phần khuôn mặt từ hình ảnh gốc.
* **Huấn luyện End-to-End mà không Cần Giám sát**: Toàn bộ hệ thống được huấn luyện theo phương thức end-to-end mà không cần giám sát cho các tọa độ điểm đặc trưng, giúp đơn giản hóa quá trình huấn luyện.

## “Vấn đề Phân tích Toàn Diện

**Phương pháp Truyền thống (CNN)**

**Cách tiếp cận**: Các phương pháp truyền thống thường xử lý hình ảnh khuôn mặt như một khối duy nhất, tức là mô hình CNN sẽ lấy toàn bộ hình ảnh khuôn mặt và trích xuất các đặc trưng từ đó. Điều này có nghĩa là mọi thông tin về các phần khác nhau của khuôn mặt (như mắt, mũi, miệng) đều được kết hợp lại mà không phân biệt rõ ràng.

**Hạn chế**: Khi khuôn mặt bị che khuất hoặc có sự biến đổi về góc nhìn, các mô hình này có thể không nhận diện chính xác vì chúng không thể tập trung vào các phần quan trọng của khuôn mặt. Ví dụ, nếu một người đeo kính râm, mắt sẽ không được nhìn thấy và mô hình CNN có thể không nhận diện được người đó.

**Phương pháp *part fViT***

**Cách tiếp cận**: Phương pháp *part fViT* giải quyết vấn đề này bằng cách sử dụng một mạng CNN nhẹ để dự đoán tọa độ của các điểm đặc trưng trên khuôn mặt, từ đó trích xuất các phần khuôn mặt riêng lẻ (như mắt, mũi, miệng). Sau đó, các phần này được đưa vào Vision Transformer (ViT) để nhận dạng.

**Lợi ích**: Bằng cách phân tích từng phần của khuôn mặt, *part fViT* có khả năng nhận diện tốt hơn ngay cả khi một số phần bị che khuất. Ví dụ, nếu chỉ có một mắt hiển thị trong hình ảnh, mô hình vẫn có thể sử dụng thông tin từ phần mũi và miệng để xác định danh tính của người đó. Điều này cho phép mô hình đạt được độ chính xác cao hơn trong các tình huống thực tế phức tạp.“

1. Đánh giá so với các pp trước, ưu nhược điểm?

**Ưu điểm:**

**Khả năng Phân tích Theo Phần:** part fViT cho phép phân tích khuôn mặt theo từng phần (như mắt, mũi, miệng) thay vì xử lý toàn bộ hình ảnh như các mô hình CNN truyền thống. Điều này giúp cải thiện độ chính xác trong các tình huống mà một số phần của khuôn mặt bị che khuất hoặc có sự biến đổi về góc nhìn1

**Không Cần Giám sát cho Tọa độ Điểm Đặc trưng:** Hệ thống được huấn luyện theo phương thức end-to-end mà không cần giám sát cho các tọa độ điểm đặc trưng. Điều này giảm thiểu độ phức tạp trong quá trình huấn luyện và tăng tính linh hoạt của mô hình1

**Sử dụng Vision Transformer:** Việc áp dụng Vision Transformer (ViT) cho phép xử lý thông tin từ các lưới không đều (irregular grids), giúp mô hình có thể hoạt động hiệu quả hơn so với các phương pháp truyền thống yêu cầu lưới đều để thực hiện các phép toán tích chập.

**Độ Chính xác Cao Hơn**: part fViT đã đạt được độ chính xác cao hơn so với nhiều phương pháp hiện tại trên các bộ dữ liệu benchmark nổi tiếng, thiết lập tiêu chuẩn mới trong lĩnh vực nhận dạng khuôn mặt.

**Khả năng Khám Phá Điểm Đặc trưng Không Giám sát:** Mạng CNN nhẹ trong pipeline không chỉ giúp dự đoán tọa độ mà còn có khả năng khám phá điểm đặc trưng mà không cần giám sát, điều này mở ra hướng đi mới cho việc phát triển các phương pháp nhận dạng khuôn mặt.

**Nhược điểm:**

* **Phụ thuộc vào Chất lượng Dự đoán Tọa độ Điểm Đặc trưng**: Phương pháp này dựa vào một mạng CNN nhẹ để dự đoán tọa độ của các điểm đặc trưng trên khuôn mặt. Nếu chất lượng dự đoán không chính xác, điều này có thể ảnh hưởng tiêu cực đến hiệu suất nhận dạng, vì các phần khuôn mặt được trích xuất từ tọa độ không chính xác có thể dẫn đến thông tin không đầy đủ hoặc sai lệch.
* **Yêu cầu về Tài nguyên Tính toán:** Các mô hình ViT thường yêu cầu tài nguyên tính toán lớn hơn so với CNN truyền thống. Điều này có thể làm cho việc triển khai trên các thiết bị có hạn chế về phần cứng trở nên khó khăn hơn.
* **Vẫn gặp khó khăn trong trong dữ liệu quy mô nhỏ.**

1. **TRANSFACE**
2. Ngữ cảnh của bài toán là gì?

Ngữ cảnh chính của bài báo là các Transformer hiện đại gặp khó khăn khi áp dụng vào bài toán nhận diện khuôn mặt trên những tập dữ liệu cực kỳ lớn. Điều này xuất phát từ việc các kỹ thuật tăng cường dữ liệu và khai thác mẫu khó hiện tại **không** **tương thích** với cấu trúc đặc biệt của ViTs, dẫn đến hiện tượng overfitting và sự phụ thuộc quá mức vào các vùng khuôn mặt nhất định (ví dụ như mắt và trán).

1. Vấn đề gì đang gặp phải trong ngũ cảnh này của các pp cũ?

Các phương pháp cũ gặp phải một số vấn đề chính khi áp dụng Vision Transformers (ViTs) vào nhiệm vụ nhận diện khuôn mặt (FR) trên các tập dữ liệu lớn:

* 1. **Overfitting**: Các phương pháp tăng cường dữ liệu hiện tại như Random Erasing, Mixup, CutMix, và RandAugment thường được sử dụng để giảm overfitting, nhưng khi áp dụng vào nhận diện khuôn mặt, chúng phá hủy một số thông tin cấu trúc quan trọng của khuôn mặt, dẫn đến việc mô hình **tối ưu hóa theo hướng sai lệch.** (Khi xóa như vậy có thể sẽ loại bỏ đi những thông tin quan trọng, mô hình có thể sẽ học những đặc trưng không lien quan hoặc học sai cách – mô hình sẽ học những đặc trưng từ vùng bị xóa thay vì học các đặc trưng khác của khuông mặt hoặc là nó chỉ có thể hoạt động trên các ảnh được thay đổi một cách tương tự, nhưng khi trong thực tế thì không).
  2. **Thiếu sự tận dụng thông tin cục bộ**: ViTs có xu hướng dựa quá nhiều vào một số mảnh khuôn mặt quan trọng (dominant patches) như mắt hoặc trán. Điều này dẫn đến việc nếu các mảnh này bị che khuất hoặc thay đổi (như khi đeo kính râm hoặc đội mũ), mô hình sẽ đưa ra dự đoán không chính xác. Các phương pháp hiện tại không khai thác đầy đủ thông tin từ các vùng khác ít được chú ý như mũi, miệng, và tai, làm giảm khả năng tổng quát hóa của mô hình.
  3. **Khai thác mẫu khó chưa hiệu quả**: Các phương pháp khai thác mẫu khó cũ (hard sample mining) như Focal Loss hoặc MV-Softmax chủ yếu dựa vào các chỉ số ở cấp độ toàn cục (global-level), chẳng hạn như xác suất dự đoán hoặc tổn thất (loss). Tuy nhiên, điều này không phù hợp với ViTs vì ViTs có thể bị chi phối bởi một số ít các token cục bộ (local tokens), dẫn đến việc đánh giá độ khó của mẫu chưa chính xác, đặc biệt khi các token quan trọng bị bỏ qua.

1. Họ đề xuất pp gì, giúp cải thiện vấn đề gì trong ngữ cảnh nào?

Để khắc phục những vấn đề này, bài báo đề xuất hai phương pháp chính:

1. **Dominant Patch Amplitude Perturbation (DPAP)**: Chiến lược tăng cường dữ liệu dựa trên mức độ ưu tiên các mảnh vá hình ảnh quan trọng, giúp tăng cường sự đa dạng của mẫu mà không làm mất thông tin cấu trúc quan trọng của khuôn mặt.
2. **Entropy-guided Hard Sample Mining (EHSM)**: Phương pháp khai thác mẫu khó dựa trên lý thuyết thông tin, giúp điều chỉnh trọng số của các mẫu dễ và khó trong quá trình huấn luyện, cải thiện độ ổn định của dự đoán.
3. Chi tiết giải pháp của họ?

### 1. ****Dominant Patch Amplitude Perturbation (DPAP)****

**DPAP** là một chiến lược tăng cường dữ liệu được thiết kế để giúp ViTs tránh bị overfitting khi xử lý các hình ảnh khuôn mặt trong bài toán nhận diện khuôn mặt (Face Recognition - FR). Mục tiêu chính của DPAP là bảo vệ các thông tin cấu trúc quan trọng của khuôn mặt, đồng thời tăng sự đa dạng của mẫu dữ liệu mà không làm mất đi những đặc điểm cần thiết.

#### Chi tiết phương pháp:

* **Vấn đề cần giải quyết**: ViTs có xu hướng học quá mức từ một vài vùng quan trọng của khuôn mặt, ví dụ như mắt và trán. Điều này khiến cho mô hình trở nên dễ bị overfitting, đặc biệt khi các vùng này bị che khuất (như khi đeo kính râm). Các phương pháp tăng cường dữ liệu hiện tại thường làm mất các thông tin cấu trúc quan trọng, dẫn đến hiệu suất giảm khi gặp những trường hợp thực tế.
* **Giải pháp của DPAP**:
  1. **Sử dụng SE module để chọn các mảnh vá quan trọng (dominant patches)**: DPAP sử dụng một module "Squeeze-and-Excitation" (SE) để xác định những mảnh vá (patch) nào của khuôn mặt đóng vai trò quan trọng nhất đối với dự đoán cuối cùng của mô hình. Các mảnh này được gọi là **dominant patches**. (Khối SE sẽ tính toán độ quan trọng của từng patch từ đó lựa chọn ta Top-K những patch chứa thông tin quan trọng nhất để phục vụ cho quá trình trộn biên độ, ngoài ra SE module còn tính toán ra một bộ trọng số để làm nổi bật lên những điểm đặc trưng và loại bỏ đi các đặc trưng không cần thiết, giúp mô hình có thể tập chung vào một số đặc trưng quan trọng hơn)
  2. **Trộn và xáo trộn thông tin biên độ của các mảnh quan trọng**: DPAP thay đổi thông tin biên độ của các dominant patches mà không thay đổi thông tin pha (phase information) của chúng. Cụ thể, DPAP áp dụng phép biến đổi Fourier FFT(Fast Fourier Transform) trên các mảnh vá, sau đó trộn thông tin biên độ của chúng với một mảnh vá ngẫu nhiên khác để tạo ra một hình ảnh mới. Tuy nhiên, thông tin pha – quan trọng để bảo tồn cấu trúc của khuôn mặt – vẫn được giữ nguyên.
  3. **Quay lại không gian hình ảnh**: Sau khi trộn xong, DPAP sử dụng phép biến đổi ngược Fourier IFFT(Inverse Fast Fourier Transform) để đưa mảnh vá về không gian hình ảnh ban đầu, tạo ra một hình ảnh khuôn mặt mới nhưng vẫn giữ được cấu trúc khuôn mặt quan trọng.
* **Tác dụng**: DPAP tạo ra những mẫu dữ liệu mới mà không làm mất đi các thông tin cấu trúc quan trọng của khuôn mặt. Điều này giúp mô hình học được nhiều hơn từ các vùng ít được chú ý (như mũi, miệng, tai) và giảm hiện tượng overfitting khi các vùng quan trọng bị thay đổi hoặc che khuất.

1. **Entropy-guided Hard Sample Mining (EHSM)**

EHSM là một chiến lược khai thác mẫu khó được thiết kế đặc biệt cho ViTs để giúp mô hình học tốt hơn từ những mẫu khó, từ đó tăng cường khả năng phân loại chính xác các mẫu dữ liệu phức tạp trong bài toán nhận diện khuôn mặt.

* **Chi tiết phương pháp**: Vấn đề cần giải quyết: Các phương pháp khai thác mẫu khó hiện tại không hoạt động tốt với ViTs vì chúng thường chỉ dựa vào các chỉ số toàn cục (global-level indicators) như xác suất dự đoán hoặc giá trị tổn thất để đánh giá độ khó của mẫu. Tuy nhiên, trong ViTs, dự đoán của mô hình có thể bị chi phối bởi một số ít các token cục bộ (local tokens), nghĩa là các chỉ số toàn cục không phản ánh đúng độ khó của mẫu.
* **Giải pháp của EHSM**:
* **Tính entropy của các token cục bộ**: EHSM đo lường độ khó của một mẫu bằng cách tính toán entropy (độ bất định thông tin) từ các token cục bộ. Token cục bộ ở đây là các đặc trưng nhỏ được rút trích từ các vùng khác nhau của khuôn mặt. Một mẫu dễ sẽ có thông tin (entropy) lớn vì nó chứa nhiều đặc điểm rõ ràng và dễ nhận biết. Ngược lại, mẫu khó (chẳng hạn như hình ảnh mờ hoặc bị che khuất) sẽ có entropy thấp vì chứa ít thông tin hữu ích.
* **Tính trọng số cho từng mẫu**: Dựa trên entropy được tính từ các token cục bộ, EHSM gán trọng số khác nhau cho các mẫu dễ và mẫu khó. Các mẫu khó (có entropy thấp) sẽ được gán trọng số cao hơn trong quá trình huấn luyện, buộc mô hình phải chú ý nhiều hơn đến những mẫu này để học các đặc trưng khó nhận diện.
* **Kết hợp với loss function**: Trọng số của mỗi mẫu (dựa trên entropy) sẽ được tích hợp vào hàm loss. Điều này giúp mô hình không chỉ học tốt từ các mẫu dễ, mà còn tận dụng được thông tin từ các mẫu khó, giúp cải thiện độ chính xác và khả năng tổng quát hóa của mô hình.

Tác dụng: EHSM giúp ViTs khai thác đầy đủ thông tin từ các vùng khác nhau trên khuôn mặt, đặc biệt là các vùng ít được chú ý (như miệng, tai, cằm). Điều này không chỉ giúp mô hình học được từ các mẫu khó, mà còn tăng cường khả năng nhận diện tổng quát khi các phần quan trọng của khuôn mặt bị che khuất.

1. Đánh giá so với các pp trước, ưu nhược điểm?

### Ưu điểm của TransFace:

1. **Hiệu suất tốt trong nhận diện khuôn mặt**:
   * TransFace đã đạt được hiệu suất cao trong các bài kiểm tra và benchmark về nhận diện khuôn mặt, vượt qua nhiều mô hình CNN truyền thống và các mô hình ViTs khác. Điều này cho thấy khả năng của nó trong việc học các đặc trưng khuôn mặt phức tạp.
2. **Chiến lược tăng cường dữ liệu hiệu quả (DPAP)**:
   * Phương pháp **Dominant Patch Amplitude Perturbation (DPAP)** giúp bảo toàn thông tin cấu trúc quan trọng trong khuôn mặt trong khi vẫn tạo ra sự đa dạng trong dữ liệu. Điều này giảm thiểu overfitting và giúp mô hình tổng quát hóa tốt hơn.
3. **Khai thác mẫu khó hiệu quả (EHSM)**:
   * Chiến lược **Entropy-guided Hard Sample Mining (EHSM)** cho phép mô hình học từ những mẫu khó, giúp cải thiện khả năng nhận diện trong các tình huống thực tế nơi mà thông tin có thể bị che khuất hoặc nhiễu.
4. **Tính khả thi và mở rộng**:
   * TransFace có khả năng mở rộng và có thể được áp dụng cho nhiều bài toán khác nhau trong lĩnh vực thị giác máy tính, không chỉ giới hạn ở nhận diện khuôn mặt.
5. **Sử dụng mô hình Transformer**:
   * Việc áp dụng cấu trúc Transformer giúp TransFace khai thác mối quan hệ giữa các đặc trưng trong ảnh một cách hiệu quả, nhờ vào cơ chế tự chú ý (self-attention)

Nhược điểm:

* 1. **Tính toán phức tạp và yêu cầu tài nguyên cao:** Mô hình Transformer nói chung và TransFace nói riêng có thể yêu cầu tài nguyên tính toán lớn và thời gian huấn luyện lâu hơn so với các mô hình CNN truyền thống. Điều này có thể là một rào cản trong việc triển khai trên các thiết bị có hạn chế về tài nguyên.
  2. **Cần nhiều dữ liệu huấn luyện:** ViTs thường yêu cầu một lượng lớn dữ liệu để hoạt động hiệu quả. Nếu dữ liệu huấn luyện không đủ phong phú, mô hình có thể không đạt được hiệu suất tối ưu, dễ dẫn đến overfitting hoặc không học được các đặc trưng quan trọng.
  3. Mặc dù DPAP và EHSM được thiết kế để cải thiện hiệu suất, nhưng nếu không được điều chỉnh đúng cách, chúng có thể dẫn đến việc mô hình học những đặc trưng không liên quan hoặc bị ảnh hưởng bởi nhiễu trong dữ liệu.

--> tiếp tục code đánh giá ACC, FPIR, FNIR trên 1 bộ dataset,

-> đánh giá trên các method CNN: Arcface, Elastic Face, AdaFace.

-> đánh giá trên các method transformer mà đã trình bày.

Swinface

* 1. Ngữ cảnh của bài toán là gì?

Bài báo nhấn mạnh rằng các mô hình transformer đã đạt được những bước tiến vượt bậc trong lĩnh vực nhận diện và phân tích gương mặt, nhưng hầu hết các phương pháp trước đó chỉ tập trung vào một nhiệm vụ cụ thể, dẫn đến việc không tận dụng được sự cộng hưởng giữa các nhiệm vụ.

* 1. Vấn đề gì đang gặp phải trong ngữ cảnh này của các pp cũ?

1. **Thiếu dữ liệu**: Các nhiệm vụ phân tích gương mặt như nhận diện biểu cảm, ước lượng tuổi và phân loại thuộc tính gương mặt thường gặp khó khăn do thiếu dữ liệu huấn luyện lớn. Điều này dẫn đến việc khi tập trung vào một nhiệm vụ duy nhất, dữ liệu lớn từ nhận diện gương mặt không thể hỗ trợ đào tạo cho các nhiệm vụ phân tích gương mặt thông qua việc chia sẻ kiến thức giữa các nhiệm vụ
2. **Hiệu quả mô hình**: Việc học các mạng riêng biệt cho các nhiệm vụ khác nhau 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 đầy đủ trong các transformer
3. **Xung đột giữa các nhiệm vụ**: Các mô hình trước đây thường chỉ tập trung vào một nhiệm vụ cụ thể, điều này dẫn đến xung đột giữa các mục tiêu của các nhiệm vụ khác nhau. Ví dụ, nhiệm vụ nhận diện gương mặt học cách trích xuất đại diện danh tính không phụ thuộc vào biểu cảm, trong khi nhiệm vụ nhận diện biểu cảm khuôn mặt lại không phụ thuộc vào danh tính cá nhân. Nếu hai subnet cụ thể cho từng nhiệm vụ này được tách ra từ lớp trên cùng, sẽ gây cản trở cho việc cải thiện hiệu suất do những mục tiêu xung đột.
   1. Họ đề xuất pp gì, giúp cải thiện vấn đề gì trong ngữ cảnh này? Chi tiết giải pháp của họ.
      * các module nào, đầu vào ra ntn, luồng chạy dữ liệu, thuật toán đề xuất chính là gì?
      * mindset tại sao họ lại làm ntn.
   2. Đánh giá so với các pp trước, ưu nhược điểm?
   3. **Ưu điểm**
      1. **Học đa nhiệm**: SwinFace được thiết kế để thực hiện đồng thời nhiều nhiệm vụ như nhận diện gương mặt, nhận diện biểu cảm khuôn mặt, ước lượng tuổi và ước lượng thuộc tính gương mặt. Điều này giúp mô hình tận dụng sự tương tác giữa các nhiệm vụ, cải thiện hiệu suất tổng thể so với các mô hình chỉ tập trung vào một nhiệm vụ đơn lẻ.
      2. **Kiến trúc chia sẻ**: Mô hình sử dụng một backbone Swin Transformer chung cho tất cả các nhiệm vụ, giúp giảm thiểu chi phí tính toán và bộ nhớ. Việc chia sẻ các đặc trưng giữa các subnet cho phép mô hình học hỏi từ dữ liệu lớn trong nhận diện gương mặt, từ đó cải thiện khả năng của các nhiệm vụ phân tích gương mặt.
      3. **Module Chú Ý Kênh Đa Cấp (MLCA)**: Mô hình tích hợp MLCA vào mỗi subnet phân tích gương mặt để giải quyết xung đột giữa các nhiệm vụ. MLCA cho phép lựa chọn các đặc trưng từ các mức độ và kênh tối ưu, giúp mô hình khai thác được cả thông tin cục bộ và toàn cục của gương mặt.
      4. **Hiệu suất vượt trội**: Các thí nghiệm cho thấy SwinFace đạt được độ chính xác 90.97% trên tập dữ liệu RAF-DB và lỗi ϵ là 0.22 trên CLAP2015. Đây là những kết quả tiên tiến nhất trong lĩnh vực nhận diện biểu cảm khuôn mặt và ước lượng tuổi.
      5. **Giảm thiểu thời gian đào tạo**: Bằng cách kết hợp nhiều nhiệm vụ vào một mô hình duy nhất, SwinFace giúp giảm thời gian đào tạo so với việc đào tạo từng mô hình riêng biệt cho mỗi nhiệm vụ.
      6. **Khả năng mở rộng**: Mô hình có khả năng mở rộng để bao gồm thêm nhiều nhiệm vụ phân tích khác trong tương lai, như định vị tư thế hoặc tái tạo 3D, mà không cần thay đổi cấu trúc cơ bản.
      7. **Cải thiện độ chính xác nhờ khởi tạo nhận diện gương mặt**: Việc sử dụng khởi tạo từ nhận diện gương mặt lớn giúp cải thiện đáng kể độ chính xác của các dự đoán trong các nhiệm vụ phân tích khác.

* 1. **Nhược điểm**

**slide**