

● Seminar Presentation

Topic:

Facial Landmark Localization

CSC14006 - Nhận dạng

Sinh viên thực hiện: Nhóm 10

21127731 - Nguyễn Trọng Tín

21127739 - Vũ Minh Phát

22127398 - Nguyễn Văn Minh Thiện

22127401 - Nguyễn Quang Thông

Giảng viên hướng dẫn:

PGS. TS. Lê Hoàng Thái

ThS. Dương Thái Bảo

ThS. Trương Tấn Khoa

Ngày 15 tháng 05 năm 2025

NỘI DUNG

01

Giới thiệu

02

Các công trình
liên quan

03

Phương pháp:
FaceXFormer

04

Thực nghiệm
& Demo

05

Kết luận

01

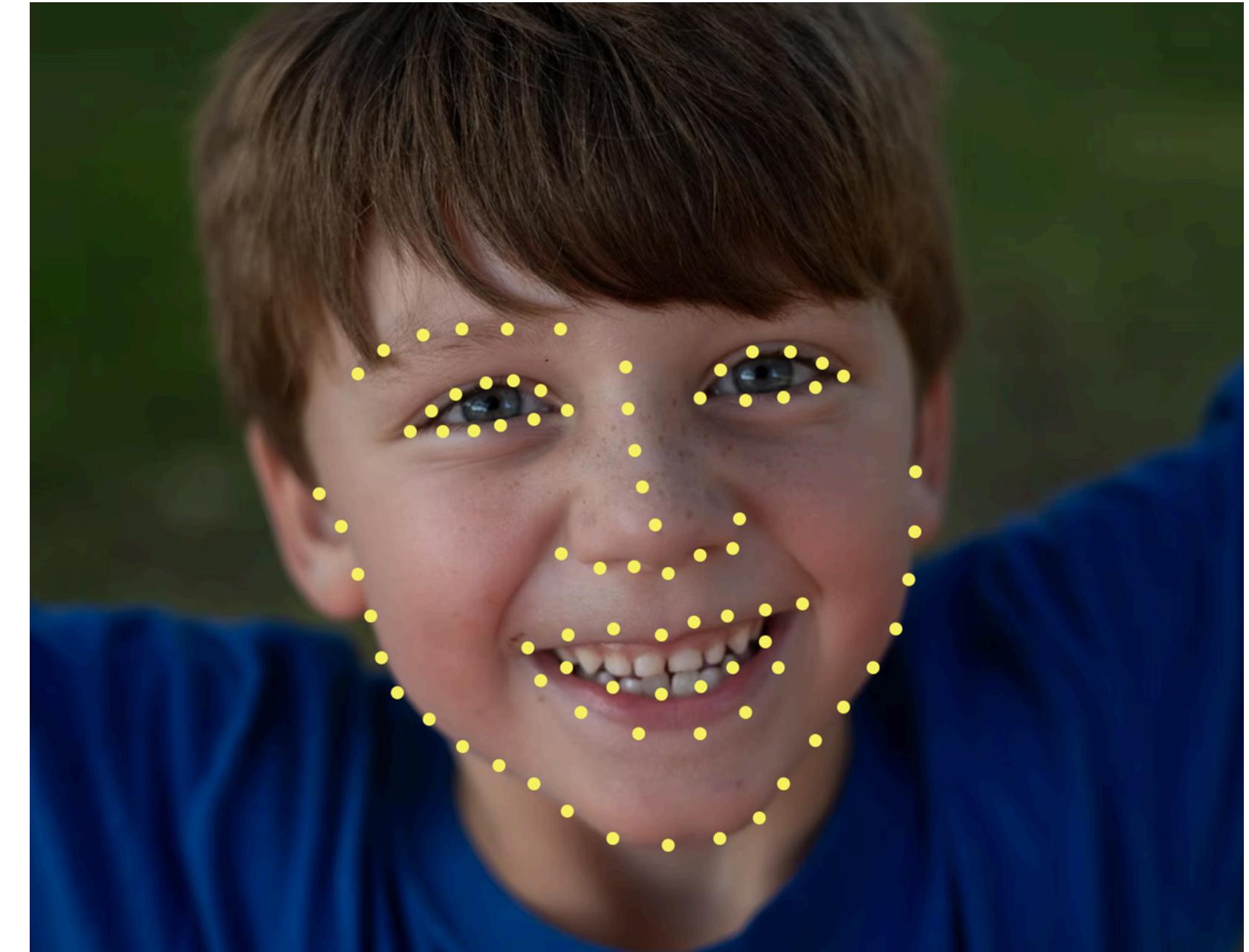
GIỚI THIỆU

Giới thiệu chủ đề: Facial Landmark Localization

Facial Landmark Localization (FLL)

Đánh dấu điểm mốc khuôn mặt (Face landmarking) là việc **phát hiện** và **xác định** vị trí các điểm đặc trưng trên khuôn mặt.

Quá trình này đóng vai trò then chốt như **một bước trung gian** cho các tác vụ xử lý khuôn mặt tiếp theo, từ *nhận dạng sinh trắc học* cho đến *phân tích trạng thái tâm lý*.

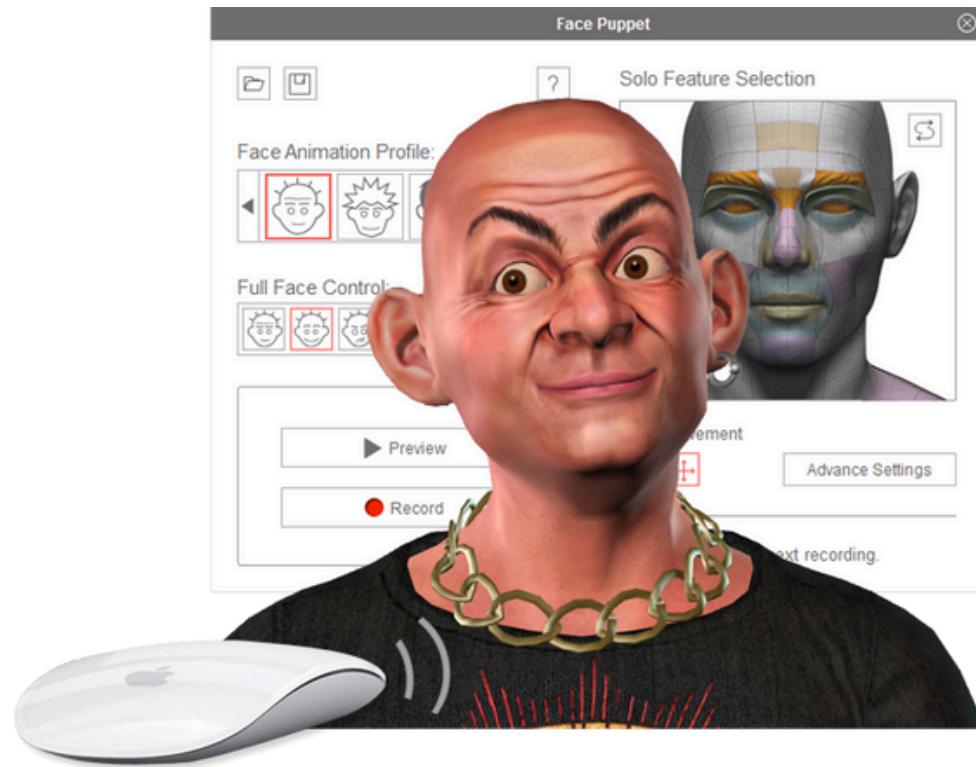


Mục tiêu là tìm ra tọa độ chính xác các điểm đặc trưng như khóm mắt, đỉnh mũi, khóm miệng, v.v..

Giới thiệu chủ đề: Facial Landmark Localization

Ứng dụng của Facial Landmark Localization

Cung cấp mô tả chi tiết về cấu trúc hình học của khuôn mặt, như là vị trí các đặc trưng trên khuôn mặt, đường viền các vùng trên khuôn mặt.



Face animation



Face Recognition



Face Editing

Giới thiệu chủ đề: Facial Landmark Localization

Thách thức của Facial Landmark Localization



Đa góc nhìn
(pose variation)



Bị che khuất
(occlusion)



Ánh sáng yếu
(poor lighting)

02

CÁC CÔNG TRÌNH LIÊN QUAN

Các phương pháp ban đầu

Đặc điểm chung

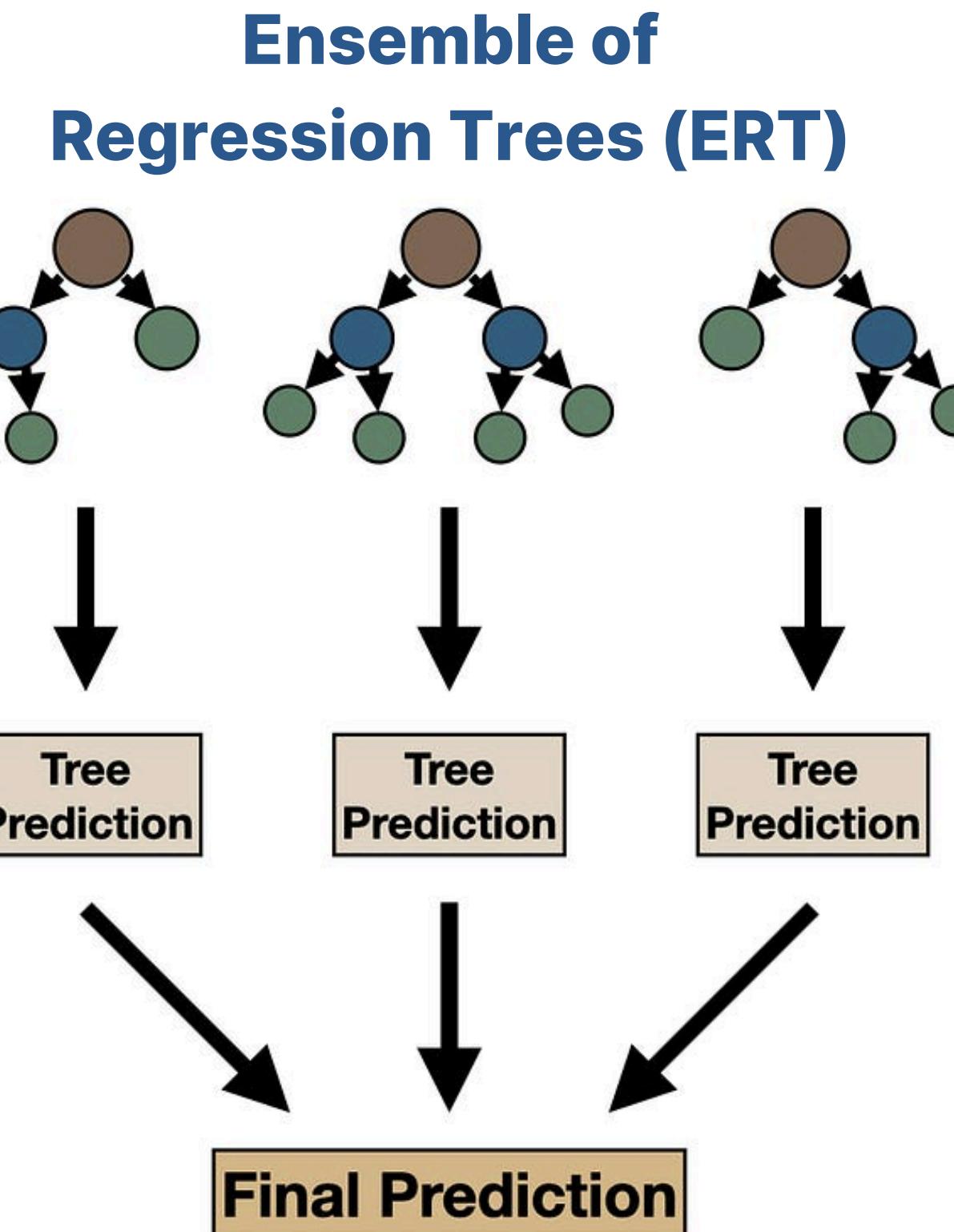
Chủ yếu dựa trên các **mô hình thống kê** và kỹ thuật học máy truyền thống

Các công trình tiêu biểu

1. Active Shape Model (ASM) - *hình dạng khuôn mặt*
2. Active Appearance Model (AAM) - *kết hợp texture*
3. Constrained Local Model (CLM)
4. **Ensemble of Regression Trees (ERT)**

Hạn chế

Khi đối mặt với các hình ảnh có *sự che khuất, tư thế lớn, hoặc điều kiện ánh sáng không lý tưởng*



Các phương pháp dựa trên mạng nơ-ron

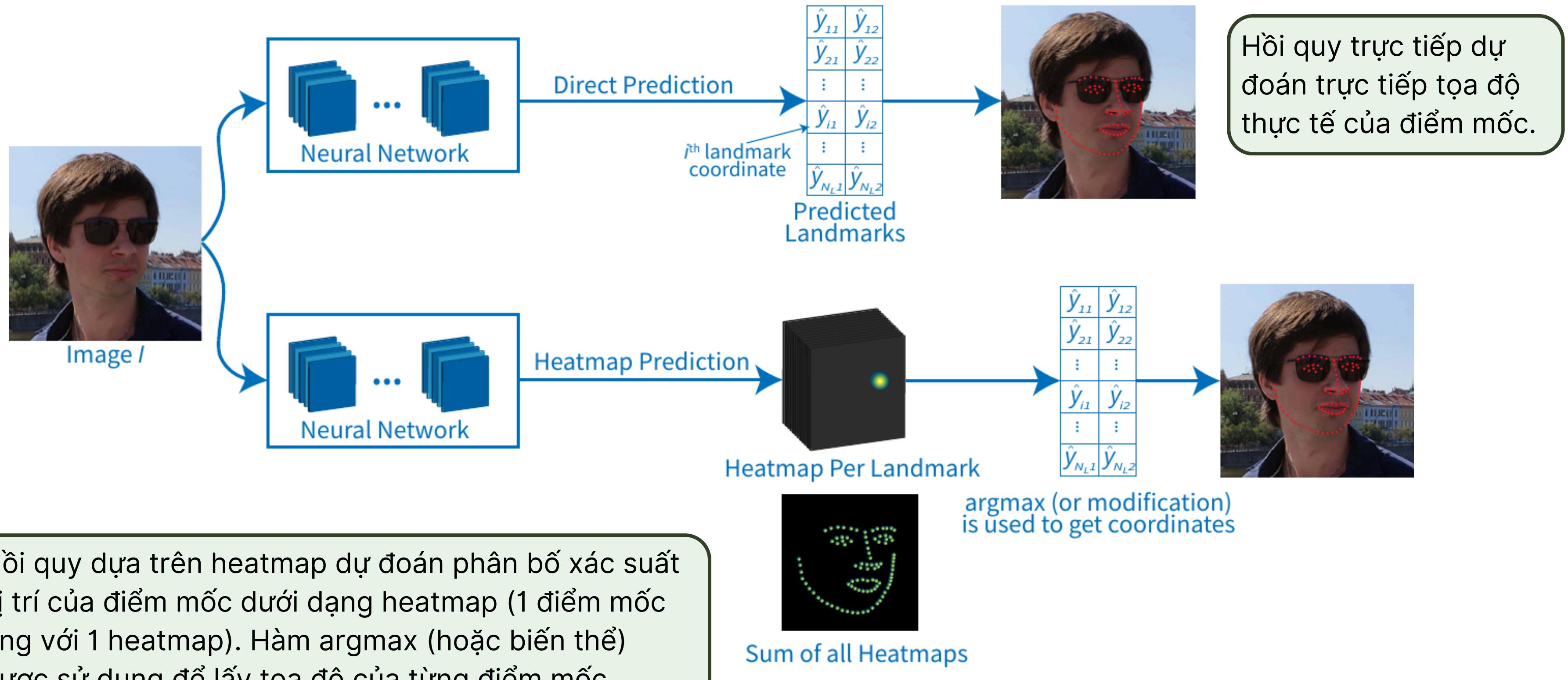
1 Động lực

Để khắc phục những hạn chế của phương pháp truyền thống, các phương pháp dựa trên **học máy, đặc biệt là học sâu**, đã được phát triển.

2 Các hướng nghiên cứu chính

- **Hồi quy trực tiếp** (direct regression)
- **Hồi quy dựa trên heatmap** (heatmap-based regression)

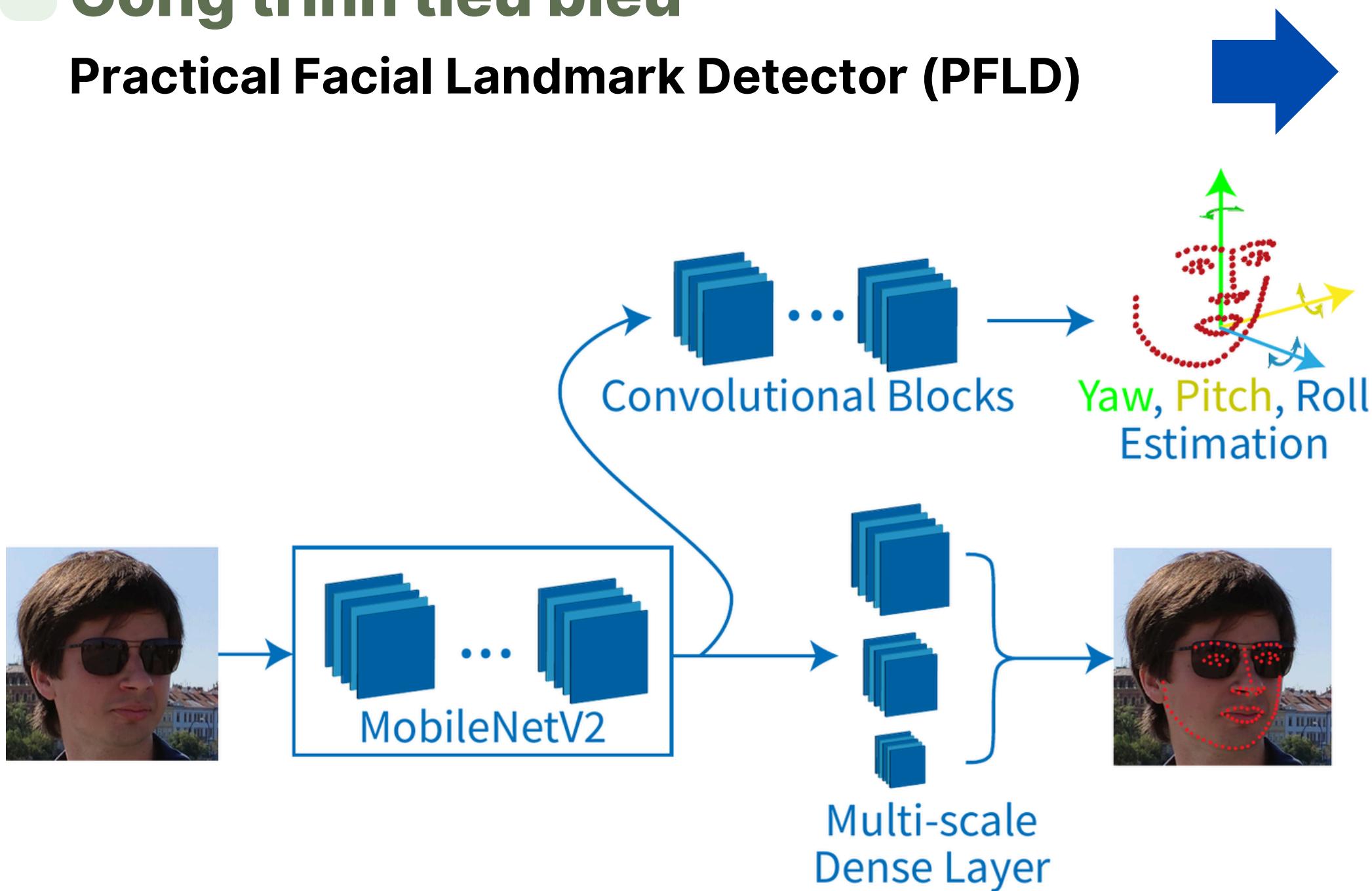
Các phương pháp dựa trên mạng nơ-ron



Phương pháp hồi quy trực tiếp

Công trình tiêu biểu

Practical Facial Landmark Detector (PFLD)



Dùng MobileNetV2 làm bộ trích xuất đặc trưng cho các tác vụ

2. Dự đoán góc quay khuôn mặt

- Sử dụng các khối tích chập bổ sung.
- Góc ước tính được dùng trong hàm mất mát khi huấn luyện để tăng hiệu suất.
- Không thực hiện dự đoán góc khi suy luận.

1. Dự đoán vị trí điểm mốc khuôn mặt

- Sử dụng lớp kết nối đầy đủ đa tỷ lệ giúp thu nhận đặc trưng ở nhiều tỷ lệ.

Phương pháp hồi quy dựa trên heatmap

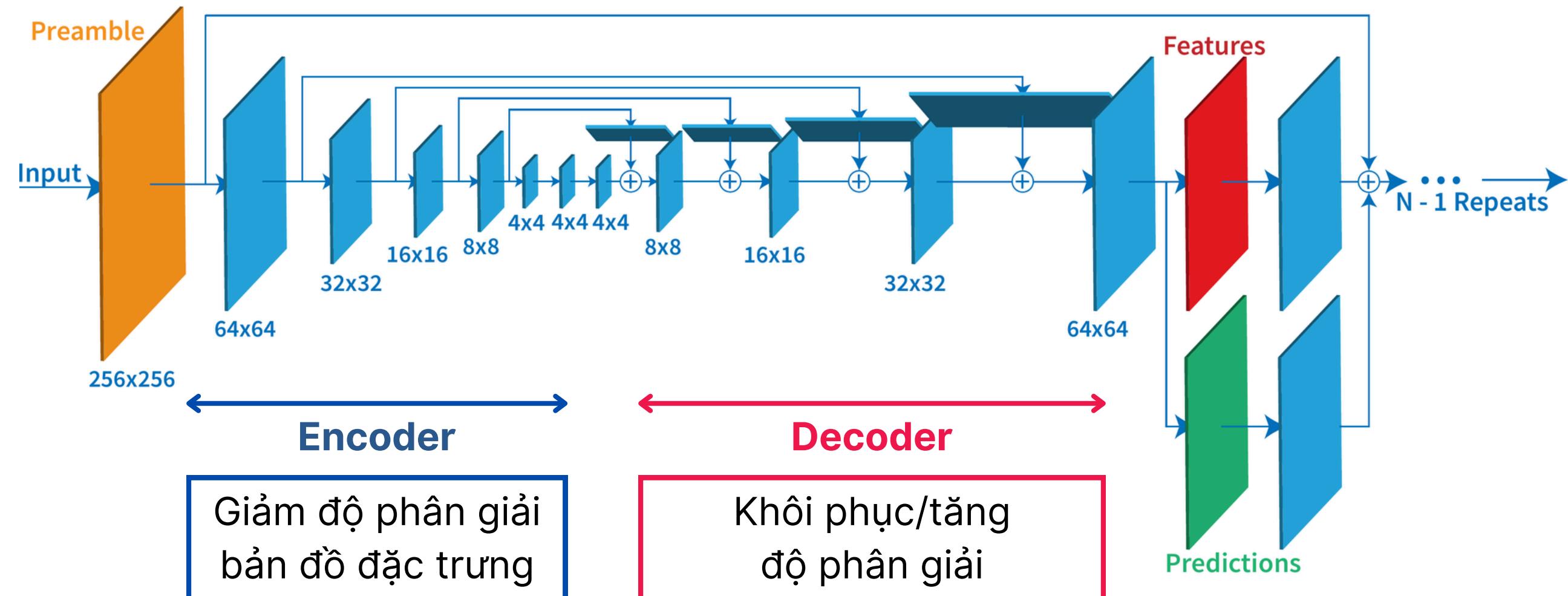
Hướng nghiên cứu

Xây dựng các backbone có khả năng **duy trì thông tin không gian ở độ phân giải cao**
→ Giúp tạo ra các heatmap chính xác

Backbone tiêu biểu

Mạng Hourglass

- Cấu trúc đối xứng:** Kiểu encoder-decoder
- Lợi ích:** Xử lý ảnh ở nhiều tỷ lệ giúp cải thiện độ chính xác cho các tác vụ



Phương pháp hồi quy dựa trên heatmap

1 Ưu điểm

- Các phương pháp dựa trên heatmap thường mang lại **độ chính xác cao** hơn so với hồi quy trực tiếp

2 Nhược điểm

- Thường **tốn kém hơn về mặt tính toán**, đặc biệt khi sử dụng các kiến trúc phức tạp như **Hourglass stack** (xếp chồng nhiều module Hourglass nối tiếp nhau)

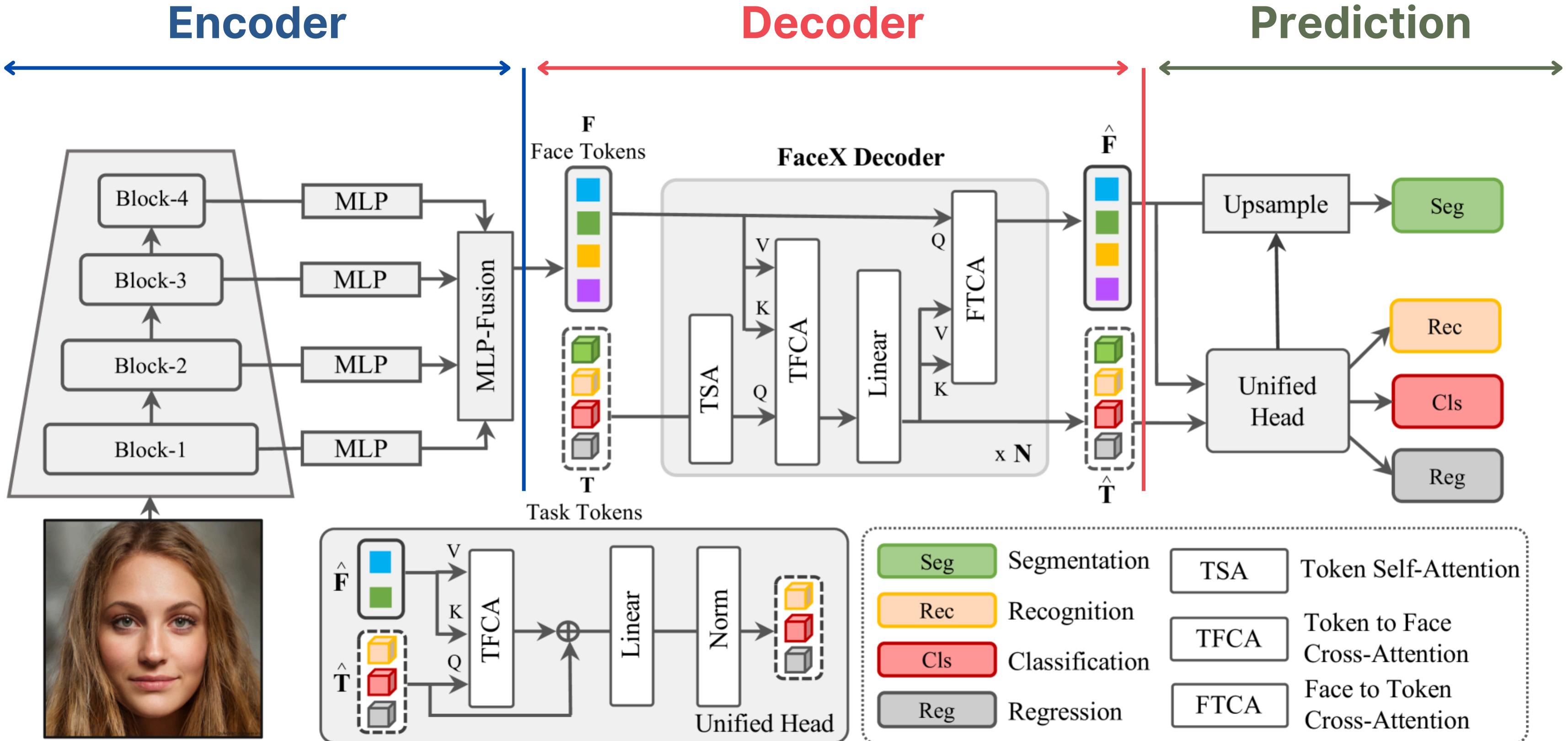
Phương pháp SOTA và kiến trúc Transformer

- 1 — **Transformer**
 - Ban đầu từ NLP, nhưng có hiệu quả cao trong Thị giác máy tính
 - **Ưu điểm:** Mô hình hóa **quan hệ tầm xa** nhờ Self-Attention
- 2 — **Xu hướng**
 - Xây dựng mô hình hợp nhất nhiều tác vụ trong một kiến trúc (multi-task learning)
- 3 — **FaceXFormer**
 - **Mô hình transformer hợp nhất**, end-to-end
 - Thực hiện đồng thời **10 tác vụ phân tích khuôn mặt**
 - Kiến trúc encoder-decoder, **mỗi tác vụ dùng một token học được**
 - Học được biểu diễn khuôn mặt tổng quát, mạnh mẽ

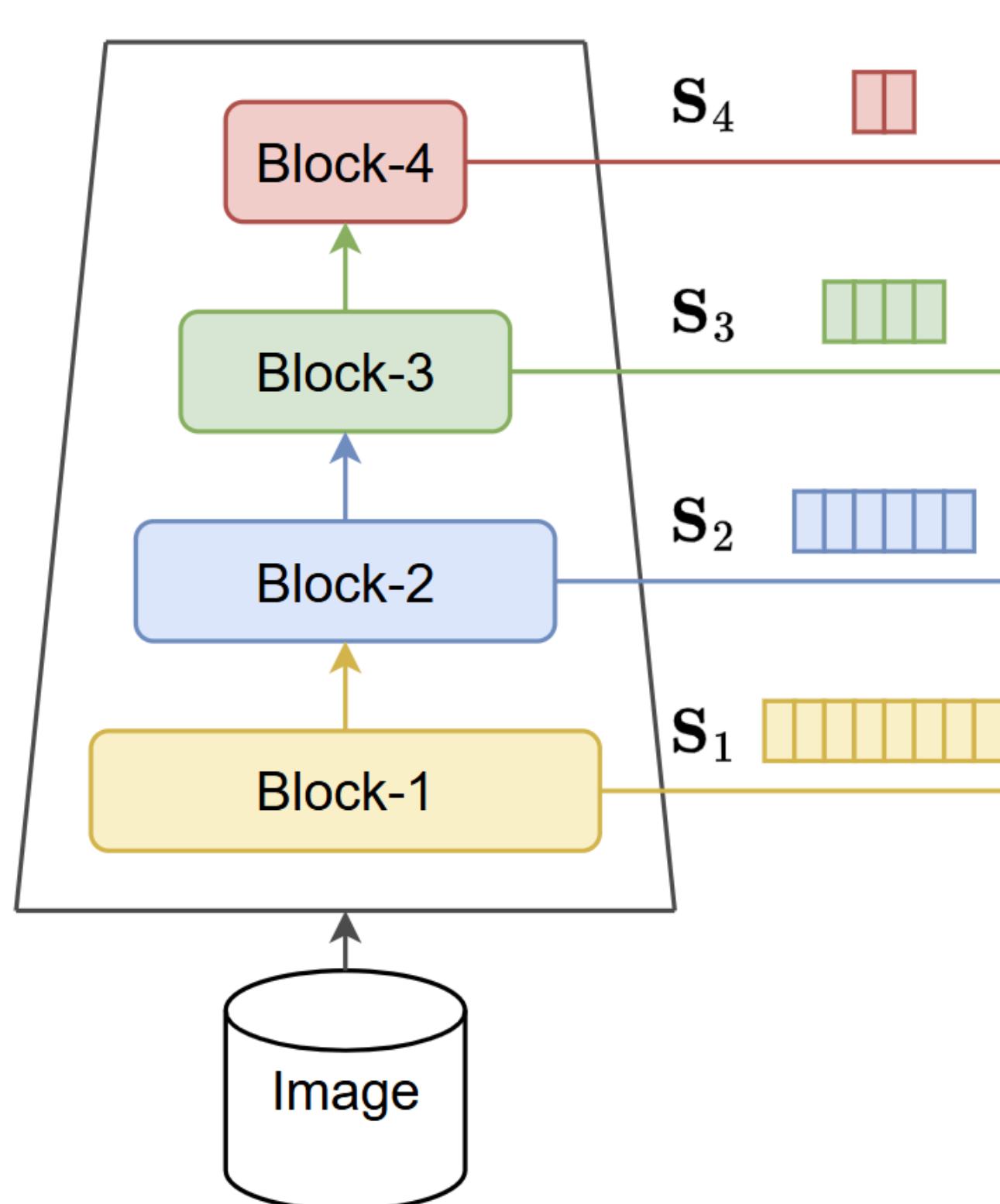
03

**PHƯƠNG PHÁP:
FACEFORMER**

Tổng quan về framework của FaceXFormer



Bộ mã hóa Đa tỷ lệ (Multi-scale Encoder)



Mục đích:

Cung cấp các loại đặc trưng khác nhau phù hợp với các tác vụ phân tích khuôn mặt đa dạng (toàn cục cho tuổi, chi tiết cho phân vùng, v.v.)

Cách hoạt động:

- Ảnh đầu vào được xử lý qua các lớp mã hóa
- Tạo ra các bản đồ đặc trưng ở nhiều mức độ trừu tượng/chi tiết khác nhau
- Đặc trưng **chuyển từ thô sơ đến chi tiết** để phù hợp với các tác vụ

Lightweight MLP-Fusion Module

1

Mục đích:

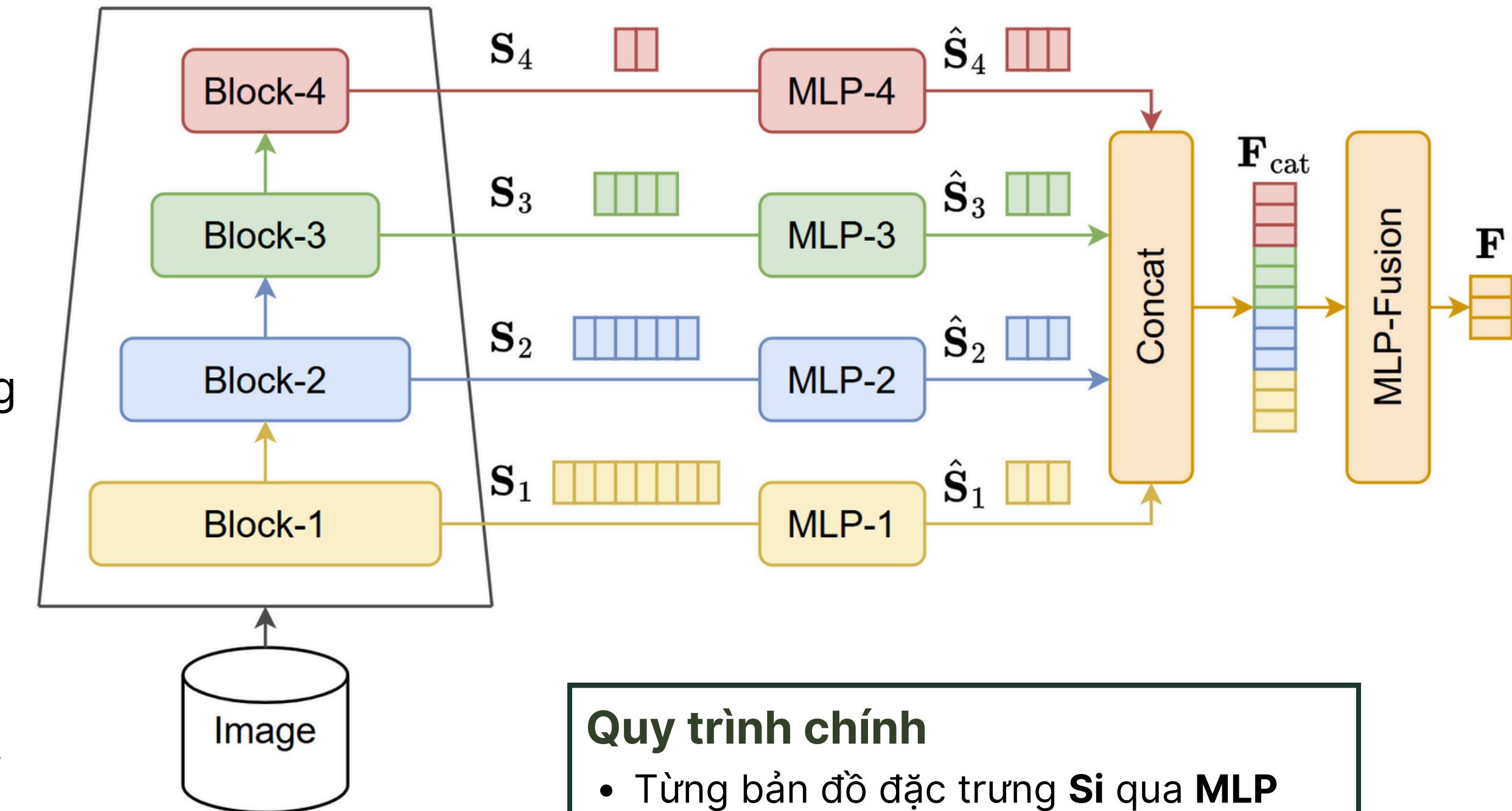
Kết hợp hiệu quả các đặc trưng đa tỷ lệ thành một biểu diễn hợp nhất

→ Hiệu quả hơn so với việc gán đặc trưng riêng cho từng tác vụ

2

Ưu điểm:

Thiết kế gọn nhẹ (~983k tham số), **chi phí tính toán thấp**, phù hợp ứng dụng thời gian thực



Quy trình chính

- Từng bản đồ đặc trưng S_i qua **MLP riêng** để **chuẩn hóa kích thước kenh**
- Ghép nối** các đặc trưng đã chuẩn hóa
- Dùng MLP hợp nhất để tạo F

FaceX Decoder (FXDec)

Mục đích

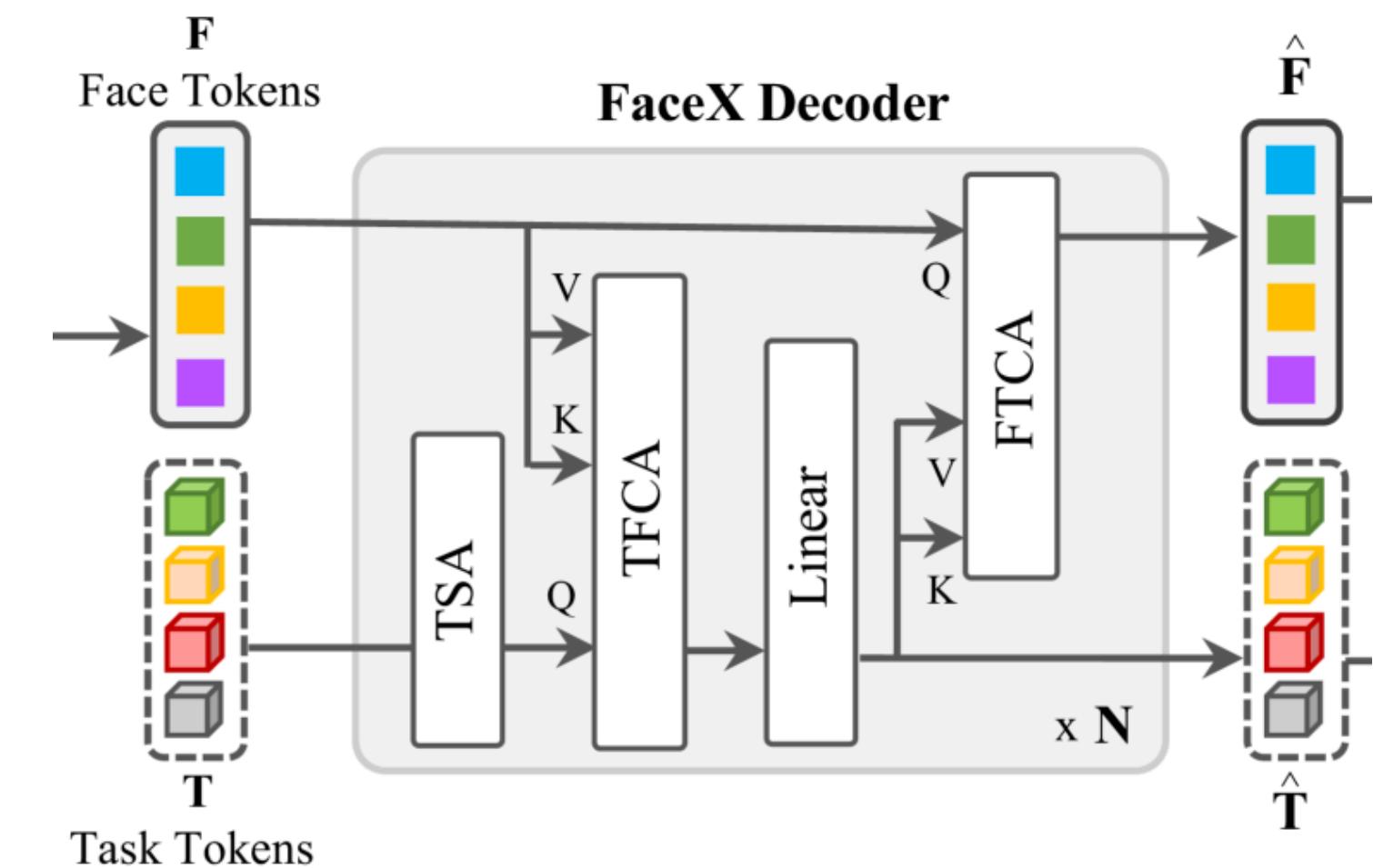
Decoder gọn nhẹ, mô hình hóa tương tác giữa Token Tác vụ và Token Khuôn mặt

→ Khắc phục nhược điểm tính toán của các Decoder truyền thống (DETR...)

Cấu trúc

Sử dụng **Cross-Attention hai chiều** gồm 3 module:

- **Task Self-Attention (TSA):** Tinh chỉnh Task Tokens, học quan hệ giữa các tác vụ (dùng **Self-Attention**).
- **Task-to-Face Cross-Attention (TFCA):** Task Tokens thu thập thông tin từ Token Khuôn mặt (**Cross-Attention: Q=Task, K/V=Face**).
- **Face-to-Task Cross-Attention (FTCA):** Token Khuôn mặt được tinh chỉnh dựa trên Task Tokens (**Cross-Attention: Q=Face, K/V=Task**).



Unified-Head (Đầu hợp nhất)

1

Mục đích:

Xử lý các token cuối cùng từ FXDec để **tạo ra dự đoán cho từng tác vụ**

2

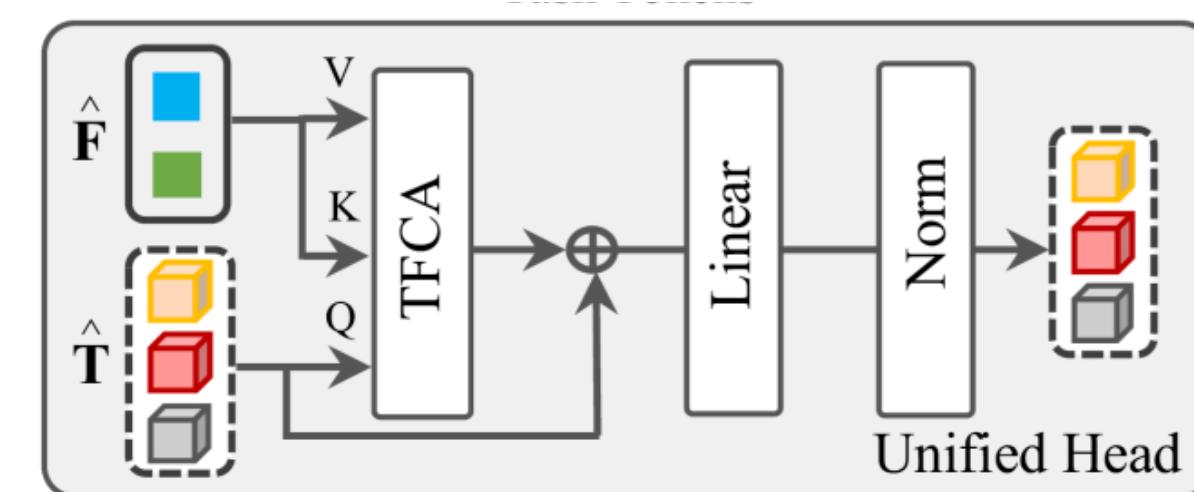
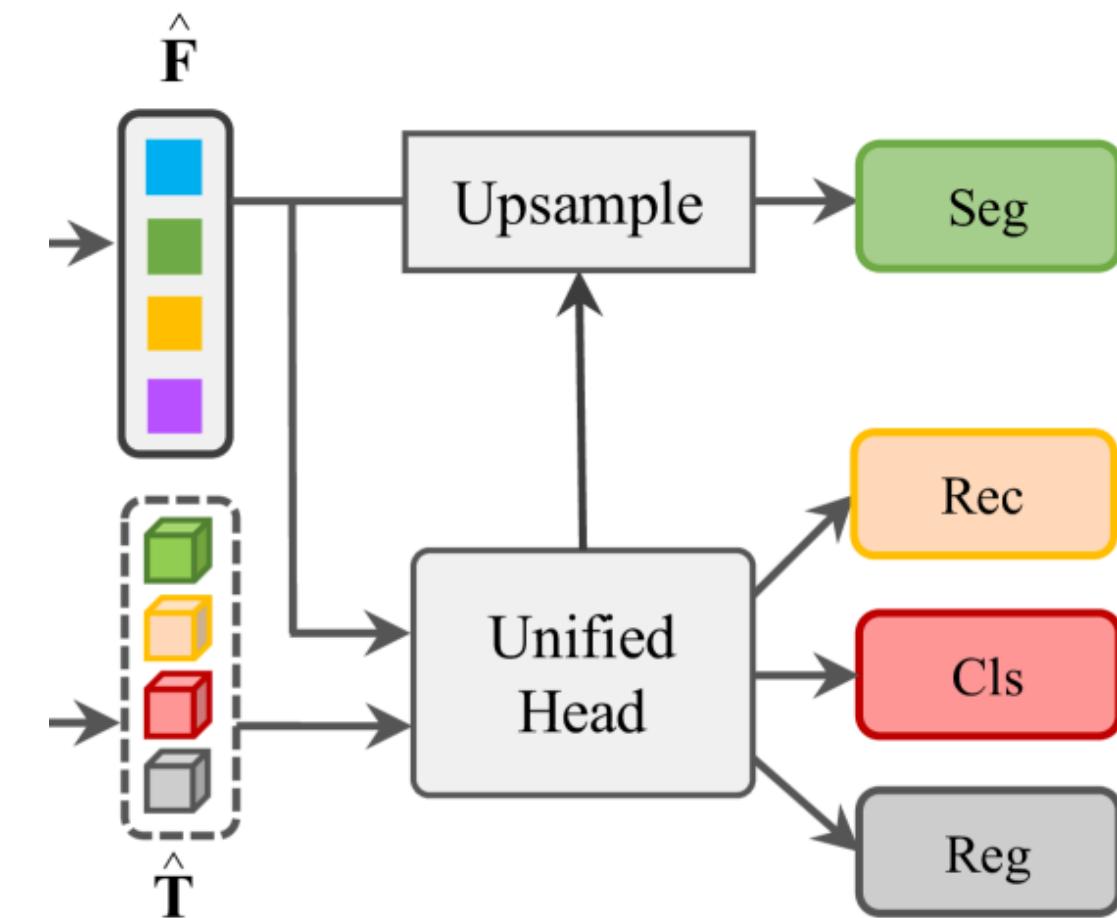
Quy trình:

- Token (\hat{F}, \hat{T}) qua module TFCA lần nữa để tinh chỉnh
- Kết quả được đưa tới các "đầu" riêng cho từng tác vụ

3

Các loại "đầu" tác vụ:

- Điểm mốc: **Mạng Hourglass** (68 token)
- Tư thế đầu: MLP hồi quy (output 9 token - ma trận xoay 3×3)
- Nhận diện: PartialFC (output 1 token)
- Tuổi, Giới tính, Chủng tộc, Biểu cảm, Khả năng hiển thị, Thuộc tính: MLP phân loại (output 1 token/tác vụ)
- Phân vùng: Upsampling F và cross-product với token phân vùng (số token = số lớp)



Huấn luyện Đa tác vụ (Multi-Task Training)

1 Cách tiếp cận

Huấn luyện đồng thời cho nhiều tác vụ phân tích khuôn mặt.

2 Mục tiêu huấn luyện

Tối ưu hóa **Hàm mất mát kết hợp (L)**

- L là **tổng có trọng số của các hàm mất mát riêng cho từng tác vụ**
- Hàm mất mát và trọng số được chọn phù hợp với từng tác vụ cụ thể

$$L = \lambda_{seg}L_{seg} + \lambda_{lnd}L_{lnd} + \lambda_{hpe}L_{hpe} + \lambda_{attr}L_{attr} + \lambda_aL_a + \lambda_{g/r}L_{g/r} + \lambda_{exp}L_{exp} + \lambda_{fr}L_{fr} + \lambda_{vis}L_{vis}$$

- *Điểm mốc:* **STAR loss**
- *Phân vùng:* Dice loss
- *Tư thế đầu:* Geodesic loss
- *Tuổi/Giới tính:* Cross-Entropy loss
- v.v..

04

THỰC NGHIỆM & DEMO

Thực nghiệm của bài báo

1 Dữ liệu sử dụng

- **300W**: 68 landmark - các góc mặt, dáng pose đa dạng.
- **AFLW**: 21 landmark - đa dạng góc quay, độ nghiêng.
- **COFW**: 29/68 landmark - các khuôn mặt bị che khuất.
- **WFLW**: 98 landmark - điều kiện khắc nghiệt nhất (pose, ánh sáng, che khuất...).

2 Đánh giá mô hình

- Chỉ số chính:
 - **NME (%)**: Sai số chuẩn hóa trung bình (càng thấp càng tốt).
 - **MAE**: Sai số trung bình tuyệt đối.
 - **FR (%)**: Tỷ lệ lỗi nghiêm trọng ($NME \geq 10\%$).
 - **AUC**: Diện tích dưới đường cong CED (càng cao càng tốt).

Thực nghiệm của bài báo

3 Thiết lập Huấn luyện

Cấu hình hệ thống

- Framework: PyTorch phân tán
- Phần cứng: 8 GPU NVIDIA A6000 (48GB)
- Input: Ảnh 224 x 224
- Backbone: Khởi tạo từ ImageNet pretrained

Tăng cường dữ liệu

- Gaussian blur, Grayscale, Gamma Adjustment
- Occlusion, Horizontal flip
- Affine Transforms: Rotate, Translate, Scale

Thông số huấn luyện

- Epoch: 12 (một số tác vụ +3 epoch)
- Batch size: 48 mẫu / GPU
- Optimizer: AdamW, Weight decay = $1e^{-5}$
- Decoder module: 2 × FaceX Decoder

Chiến lược cân bằng dữ liệu

Dùng upsampling để cân bằng các tập dữ liệu nhỏ hơn → Đảm bảo mỗi tác vụ được đại diện công bằng trong batch.

Thực nghiệm của bài báo

4

Kết quả đa nhiệm của FaceXFormer

- **10** tác vụ xử lý khuôn mặt – hiệu suất hàng đầu ở hầu hết các bài toán.
- Ví dụ nổi bật:
 - Face Parsing: F1 = 92.01%.
 - Head Pose: MAE = 3.52.
 - Landmark Detection: NME = 4.67.
 - Age Estimation: MAE = 4.17.
 - Expression Recognition: Accuracy = 88.24%.

5

Đánh giá định tính

- Hoạt động tốt trên ảnh thực tế (góc lệch, che khuất, mờ).
- Khả năng phân tích chính xác nhiều thuộc tính từ 1 ảnh:
 - Pose, landmark, tuổi, giới tính, chủng tộc, biểu cảm.
- Thích hợp cho gán nhãn tự động, giám sát, phân tích video.

Thực nghiệm của nhóm

1

Dữ liệu sử dụng

- iBUG 300W: gồm 7674 ảnh khác nhau được chia làm 3 tập khác nhau
 - Training Set: 6000 ảnh
 - Validation Set: 666 ảnh
 - Test Set: 1008 ảnh
- Mỗi ảnh đều được chú thích với 68 điểm landmarks khác nhau.
- Seed khởi tạo cũng được giữ cố định để nhất quán trong quá trình thực nghiệm.

2

Mục tiêu thực nghiệm

- Coordinate Regression (Tọa độ trực tiếp): Sử dụng WingLOSS
- Heatmap-Based: Sử dụng StarLOSS
- Kiểm tra trên Pretrained Model - FACEFORMER

Thực nghiệm của nhóm

3

Tăng cường dữ liệu

- Tăng cường sự đa dạng của tập dữ liệu huấn luyện
- Áp dụng các phương pháp với **xác suất khác nhau** như:
 - **Biến đổi hình học:** Resize, Random Affine, Random Horizontal Flip
 - **Biến đổi màu sắc:** Random GrayScale, Random Blur, Gamma Adjustement
 - **Mô phỏng che khuất:** Random Erasing

4

Thông số huấn luyện

- Số lượng Epoch: 12 epochs
- Batch Size: 48
- Learning Rate khởi tạo: 1e-4
- Optimizer: AdamW -Weight Decay: 1e-5
- Sử dụng MultiStepLR để giảm lr tại các epoch 6 và 10

Thực nghiệm của nhóm

5 Khởi tạo mô hình

- **Backbone:** Swin-B Transformer (Swin Base)
- Sử dụng **Xavier Uniform** để cân bằng các lớp gradient, hạn chế gradient vanishing/exploding

6 Hàm mất mát

Coordinate Regression Training

- WingLoss
- Các tọa độ (x, y) trực tiếp
- Chấp nhận sai số nhỏ, giảm ảnh hưởng từ outliers
- Đơn giản, tốc độ nhanh

Heatmap-Based Training

- StarLoss
- Heatmap phân bố xác suất của từng điểm
- Kết hợp các kiểm định thống kê để xử lý các vùng không gian thêm.
- Chính xác hơn nhưng tốn tài nguyên

Thực nghiệm của nhóm

7

Tối ưu hóa quá trình huấn luyện

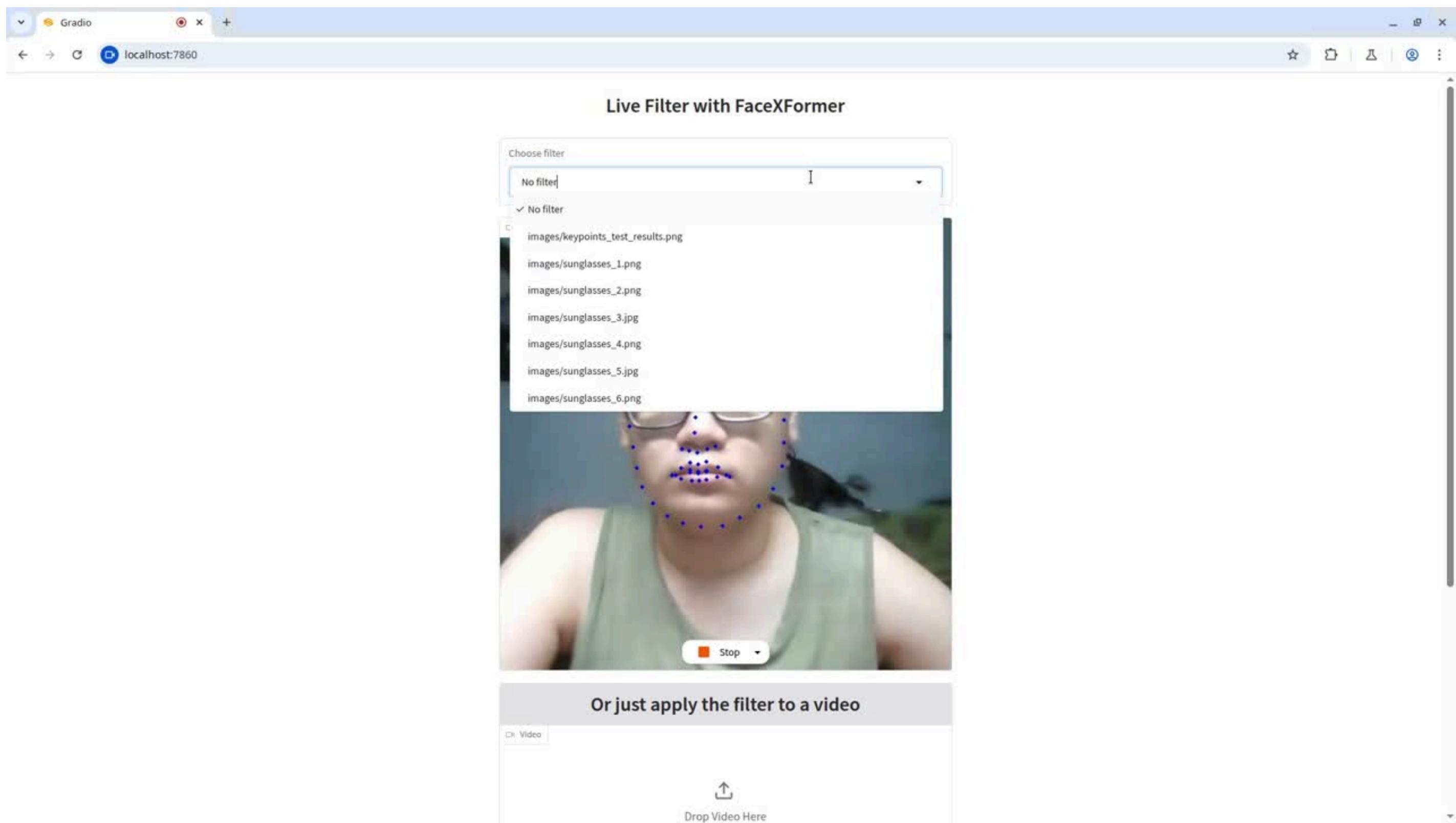
- **Mixed Precision Training:** Autocast dữ liệu giữa float 16/float 32

8

Kết quả thực nghiệm

Model	NME
Pretrained FaceXFormer	4.3664
FaceXFormer + SwinB + Coord + WingLoss	4.2830
FaceXFormer + SwinB + Heatmap + STARLoss	4.3846

Demo



05

KẾT LUẬN

KẾT LUẬN

Những thành tựu đạt được

- Là **thành phần cốt lõi** trong nhiều hệ thống: nhận diện khuôn mặt, AR, y học.
- Deep learning vượt trội phương pháp truyền thống về độ chính xác và tính khai quát.

Các thách thức

- **Tính đa dạng của dữ liệu đầu vào:** Sự thay đổi về tư thế, ánh sáng, biểu cảm, độ tuổi và sắc tộc đòi hỏi mô hình có tính khai quát cao và khả năng học biểu diễn phức tạp.
- Tình huống **che khuất** và **biến dạng**: Các trường hợp như đeo khẩu trang, quay góc lớn gây mất mốc đặc trưng, yêu cầu mô hình có khả năng suy luận theo ngữ cảnh.

Hướng nghiên cứu tương lai

- **Tích hợp Facial Landmark Localization** với các nhiệm vụ liên ngành (nhận diện cảm xúc, theo dõi chức năng cơ mặt, chẩn đoán lâm sàng) để tăng giá trị ứng dụng.
- Nghiên cứu **tối ưu mô hình** nhẹ và hiệu quả, phù hợp với thiết bị có tài nguyên hạn chế mà vẫn đảm bảo độ chính xác và độ tin cậy.

● Seminar Presentation

THANK
YOU

Ngày 15 tháng 05 năm 2025

