TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CS331.P21 - Thị giác máy tính nâng cao

Swin-SE-ResNet: Nhận diện cảm xúc khuôn mặt

Trương Huỳnh Thúy An	Hoàng Đức Chung	Nguyễn Hải Đăng
22520033	22520161	22520189

Tóm tắt nội dung

Transformer, đặc biệt là Swin Transformer, có khả năng nắm bắt thông tin ngữ cảnh toàn cục, giúp nhận diện các sự khác biệt tinh tế trong biểu cảm khuôn mặt. Tuy nhiên, các mô hình này có độ phức tạp tính toán cao và dễ overfitting nếu thiếu dữ liệu huấn luyện phong phú. Trong nghiên cứu này, chúng tôi giải quyết bài toán nhận diện cảm xúc khuôn mặt – một bài toán thị giác máy tính với nhiều ứng dụng như tương tác người-máy và phân tích hành vi. Chúng tôi đề xuất một kiến trúc lai mang tên Swin-SE-ResNet, kết hợp giữa Swin Transformer – nổi bật với khả năng khai thác ngữ cảnh toàn cục – và ResNet-50 – có khả năng trích xuất đặc trưng cục bộ mạnh mẽ. Để tăng cường khả năng biểu diễn, mô hình tích hợp thêm các mô-đun như Squeeze-and-Excitation (SE) để điều chỉnh trọng số theo kênh, Split Convolution để tăng tính đa dạng không gian, và Mean Pooling để giảm nhiễu và kiểm soát độ phức tạp. Chúng tôi khai thác song song hai tập dữ liệu FER2013 và AffectNet nhằm tăng độ đa dạng và tính tổng quát, đồng thời áp dụng kỹ thuật tiền xử lý và tăng cường dữ liệu để khắc phục mất cân bằng lớp. Kết quả thực nghiệm cho thấy mô hình đạt độ chính xác và F1-score cao, ổn định. Cụ thể, trên AffectNet, mô hình đạt độ chính xác 76.72% và F1-score 76.81%; trên FER2013 là 71.64% và 71.71%. Khi kết hợp hai tập dữ liệu, kết quả đạt 73.26% độ chính xác và 73.24% F1-score, cho thấy tính tổng quát tốt hơn.

1 Giới thiệu

Facial Expression Recognition (FER) là một bài toán quan trọng trong lĩnh vực thị giác máy tính, với nhiều ứng dụng như tương tác người–máy, phân tích hành vi người dùng, hỗ trợ điều trị tâm lý và chăm sóc sức khỏe. Tuy nhiên, bài toán này vẫn đầy thách thức, đặc biệt trong môi trường không kiểm soát (in-the-wild) khi đối mặt với các yếu tố như thay đổi ánh sáng, tư thế khuôn mặt, che khuất và sự đa dạng về chủng tộc. Các mô hình học sâu, đặc biệt là Swin Transformer, một dạng Vision Transformer có khả năng khai thác đặc trưng toàn cục – đã đem lại nhiều tiến bộ trong FER. Tuy vậy, Swin Transformer vẫn còn hạn chế trong việc học đặc trưng cục bộ chi tiết, vốn rất quan trọng để phân biệt các biểu cảm tương đồng. Ngoài ra, việc huấn luyện trên dữ liệu đơn lẻ thường dẫn đến overfitting và kém tổng quát trong thực tế.

Trong nghiên cứu này, chúng tôi đề xuất một kiến trúc kết hợp mới có tên **Swin-SE-ResNet**, tận dụng đồng thời khả năng trích xuất đặc trưng toàn cục của Swin Transformer và đặc trưng cục bộ sâu của ResNet-50. Bên cạnh đó, chúng tôi tích hợp thêm các thành phần như:

- **Squeeze-and-Excitation (SE) Block**: Tăng cường khả năng học kênh đặc trưng quan trọng;
- **Split Convolution**: Tăng tính đa dạng không gian và giảm hiện tượng đồng bộ hóa không mong muốn;
- **Mean Pooling**: Giảm nhiễu và đơn giản hóa cấu trúc mô hình.

Chúng tôi sử dụng hai bộ dữ liệu lớn là FER2013 và AffectNet, đồng thời kết hợp chúng để tăng tính đa dạng. Kỹ thuật tiền xử lý và tăng cường dữ liệu cũng được áp dụng nhằm giải quyết vấn đề mất cân bằng lớp và overfitting.

Về kết quả, mô hình Swin-SE-ResNet đạt độ chính xác khá ấn tượng, cao nhất là 76.72% và F1-score 76.81% trên AffectNet và một số kết quả chi tiết khác được

thể hiện rõ ở phần thực nghiệm 5.2. Chúng tôi cũng so sánh với các mô hình hiện đại như TransFER [14], SwinT-SE-SAM [15] và Swin-FER [5]. Kết quả thực nghiệm cho thấy mô hình của chúng tôi vượt trội hơn cả về độ chính xác và khả năng tổng quát, được trình bày chi tiết trong bảng 7 và 8.

Mục tiêu của nghiên cứu là xây dựng một mô hình nhận diện cảm xúc khuôn mặt có độ chính xác cao và khả năng tổng quát tốt trong môi trường thực tế, thông qua việc kết hợp các đặc trưng toàn cục và cục bộ của ảnh khuôn mặt. Bên cạnh đó, các đóng góp chính của đề tài bao gồm:

- Đề xuất kiến trúc lai **Swin-SE-ResNet** kết hợp *Swin Transformer* và *ResNet-50*, tích hợp thêm các thành phần như *SE-block*, *Split Convolution* và *Mean Pooling* nhằm tăng khả năng biểu diễn và giảm hiện tượng quá khớp (overfitting).
- Thiết lập quy trình huấn luyện hiệu quả với các kỹ thuật tiền xử lý và tăng cường dữ liệu nhằm khắc phục vấn đề mất cân bằng lớp và nâng cao hiệu suất mô hình.
- Triển khai và đánh giá mô hình trên ba tập dữ liệu: **FER2013**, **AffectNet** và tập dữ liệu kết hợp, sử dụng các chỉ số đánh giá đầy đủ như *accuracy*, *F1-score* và *confusion matrix*.
- Ứng dụng mô hình vào hệ thống *real-time* sử dụng webcam, nhằm kiểm chứng khả năng hoạt động trong môi trường thực tế.

2 Các nghiên cứu liên quan

Nhận diện biểu cảm khuôn mặt (FER) là một bài toán nhận được nhiều sự quan tâm trong cộng đồng nghiên cứu học sâu. Transformer, với cơ chế tự chú ý (self-attention), đã chứng minh hiệu quả cao trong bài toán FER [13]. Xue và cộng sự [14] phát triển mô hình TransFER bằng cách áp dụng Multi-Attention Dropping (MAD) để tập trung vào các vùng quan trọng và giảm nhiễu, giúp mô

hình phân biệt tốt hơn giữa các biểu cảm tương đồng. Ma và cộng sự [12] giới thiệu VTFF – một mô hình kết hợp giữa CNN hai nhánh và Transformer với cơ chế Attentional Selective Fusion (ASF), nhằm trích xuất các "từ thị giác" (visual words) phân biệt cao từ đặc trưng toàn cục và cục bộ.

Trong những năm gần đây, kiến trúc Transformer, đặc biệt là Swin Transformer, đã được ứng dụng thành công vào bài toán FER nhờ khả năng khai thác đặc trưng toàn cục mạnh mẽ thông qua cơ chế self-attention theo cửa sổ trượt. Liang và cộng sự [4] đề xuất mạng CT-DBN kết hợp CNN và Swin Transformer để giải quyết các thách thức như che khuất và thay đổi góc nhìn. Một số nghiên cứu khác cũng đã tích hợp các mô-đun attention đa cấp vào Swin Transformer nhằm chọn lọc đặc trưng hiệu quả hơn và nâng cao độ chính xác phân loại. Trong cùng năm, Vats và Chadha [15] đã đề xuất mô hình kết hợp giữa Swin Transformer và các khối Squeeze-and-Excitation (SE) nhằm tăng khả năng biểu diễn đặc trưng theo chiều sâu kênh. Ngoài SE, mô hình còn bổ sung mô-đun Spatial Attention Module (SAM) nhằm tăng cường khả năng nhận diện cảm xúc với số lượng dữ liệu vừa phải. Nhóm tác giả sử dụng tập dữ liệu tổng hợp từ FER2013 (40.000 ảnh), CK+ (1.000 ảnh) và AffectNet (60.000 ảnh), áp dụng nhiều kỹ thuật tiền xử lý như chuẩn hóa ảnh, xoay ảnh, thay đổi độ tương phản và tăng cường dữ liệu. Kết quả thực nghiệm trên AffectNet (4.000 ảnh kiểm tra) cho thấy mô hình SwinT-SE-SAM đạt F1-score cao nhất là 0.5420 – vượt trội so với các phiên bản gốc.

Dựa trên kiến trúc Swin Transformer, Bie và cộng sự [5] phát triển mô hình Swin-FER với nhiều cải tiến như hợp nhất đặc trưng giữa các tầng, sử dụng group convolution, và thêm các mô-đun như Mean/Split Module. Mô hình này đạt độ chính xác 71.11% trên FER2013 và 100% trên CK+, thể hiện hiệu quả cao trong nhiều điều kiện khác nhau.

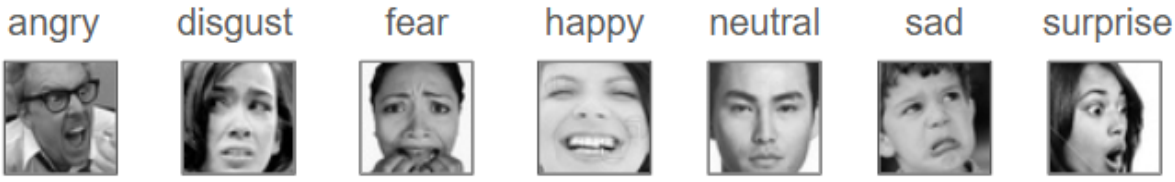
Tuy nhiên, các nghiên cứu trên vẫn còn hạn chế trong việc khai thác triệt để các đặc trưng cục bộ. Do đó, chúng tôi đề xuất một mô hình lai có tên **Swin-SE-ResNet**, kết hợp giữa Swin Transformer và ResNet-50 nhằm đồng thời học được cả đặc trưng toàn cục và cục bộ một cách hiệu quả. Bên cạnh đó, chúng tôi tích hợp khối Squeeze-and-Excitation (SE) [6] để tăng cường khả năng biểu diễn đặc

trưng theo chiều sâu kênh, đồng thời sử dụng Split Convolution và Mean Pooling nhằm giảm nhiễu và kiểm soát số lượng tham số. Mô hình được huấn luyện và đánh giá trên tập dữ liệu FER2013 và AffectNet để tận dụng sự đa dạng và nâng cao khả năng khái quát hóa — yếu tố mà các nghiên cứu trước đây chưa khai thác một cách đầy đủ.

3 Bộ dữ liệu

Trong đề tài này, chúng tôi sử dụng hai tập dữ liệu: FER2013 và AffectNet. Cả hai tập đều bao gồm các ảnh khuôn mặt với nhãn tương ứng thuộc một trong bảy cảm xúc cơ bản: Angry (giận dữ), Disgust (ghê tởm), Fear (sợ hãi), Happy (vui vẻ), Neutral (bình thường), Sad (buồn bã) và Surprise (ngạc nhiên).

FER2013 là một tập dữ liệu phổ biến được giới thiệu trong cuộc thi ICML Challenge năm 2013 [1]. Tập dữ liệu này bao gồm **35,887** ảnh xám với kích thước 48×48 pixel, được thu thập trong môi trường thực tế và được gán nhãn cảm xúc theo 7 lớp với một vài ảnh minh họa trong hình 2. Và nó không đồng đều về phân bố các lớp cảm xúc.



Hình 1: Một số hình ảnh mẫu từ tập dữ liệu Fer2013.

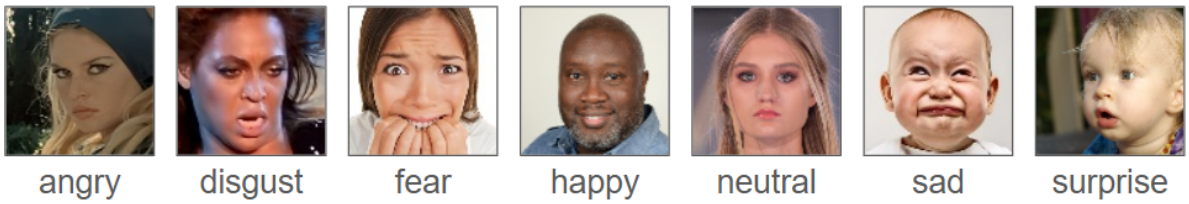
Bảng 1 cho thấy sự mất cân bằng rõ rệt giữa các lớp cảm xúc trong tập dữ liệu FER2013. Cảm xúc Happy chiếm tỷ lệ lớn nhất với 8,989 ảnh, gần gấp đôi so với nhiều lớp khác. Trong khi đó, lớp Disgust có số lượng thấp nhất với chỉ 547 ảnh, tạo nên sự chênh lệch lớn so với các lớp còn lại. Các lớp như Fear, Neutral, Sad và Angry có số lượng tương đối đồng đều, dao động từ khoảng 4,900 đến 6,200 ảnh. Sự phân bố không đồng đều này có thể gây ra hiện tượng mất cân bằng lớp, ảnh

hưởng tiêu cực đến khả năng học và tổng quát hóa của mô hình nếu không được xử lý bằng các kỹ thuật như oversampling, augmentation hoặc loss weighting.

Nhãn	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Tổng cộng	4953	547	5121	8989	6198	6077	4002

Bảng 1: Phân bố số lượng ảnh theo nhãn cảm xúc trong tập dữ liệu FER2013

AffectNet là một trong những tập dữ liệu nhận diện cảm xúc lớn và được xây dựng từ các ảnh thu thập qua internet bằng cách sử dụng các từ khóa liên quan đến cảm xúc. AffectNet chứa hơn 1 triệu ảnh, trong đó khoảng 440,000 ảnh được gán nhãn thủ công với các cảm xúc cơ bản [2]. Trong nghiên cứu này, chúng tôi lựa chọn một phần của AffectNet với tổng số lượng ảnh là **22,244** tương ứng với 7 cảm xúc giống như FER2013 với kích thước 96×96 pixel.



Hình 2: Một số hình ảnh mẫu từ tập dữ liệu AffectNet.

Nhãn	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Tổng cộng	3434	3233	2961	3325	2374	2793	4124

Bảng 2: Phân bố số lượng ảnh theo nhãn cảm xúc trong tập dữ liệu AffectNet

Bảng 2 cho thấy phân bố số lượng ảnh theo nhãn cảm xúc trong tập dữ liệu AffectNet tương đối đồng đều hơn so với FER2013, tuy nhiên hiện tượng mất cân bằng lớp vẫn tồn tại. Lớp Surprise có số lượng ảnh cao nhất với 4,124 ảnh, trong khi lớp Neutral thấp nhất với 2,374 ảnh. Các lớp còn lại dao động trong khoảng từ 2,700 đến 3,400 ảnh. Sự chênh lệch này tuy không quá lớn nhưng vẫn có thể ảnh hưởng đến khả năng học của mô hình, đặc biệt ở những lớp có ít mẫu hơn.

Không dừng lại ở việc sử dụng riêng lẻ từng tập dữ liệu, chúng tôi tiến hành kết hợp có chọn lọc hai tập FER2013 và AffectNet nhằm xây dựng một tập dữ liệu kết hợp, đa dạng và cân bằng hơn giữa các lớp cảm xúc. Cụ thể, các ảnh thuộc cùng một nhãn cảm xúc từ hai tập được gộp lại, đồng thời áp dụng kỹ thuật lọc và trích mẫu để đảm bảo phân bố đồng đều giữa các lớp với tổng số lượng ảnh thu được là là **38,550**. Việc kết hợp này không chỉ giúp tăng kích thước tổng thể của tập dữ liệu huấn luyện mà còn góp phần giảm thiểu tình trạng mất cân bằng lớp, một yếu tố thường làm suy giảm hiệu quả học của mô hình phân loại.

Nhãn	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
Tổng cộng	5992	3900	5589	5568	5894	5661	5946

Bảng 3: Phân bố số lượng ảnh theo nhãn cảm xúc trong tập dữ liệu kết hợp (FER2013 + AffectNet)

Trước khi huấn luyện mô hình, toàn bộ dữ liệu ảnh trong ba tập **huấn luyện**, **kiểm tra** và **kiểm định** đều được chuẩn hóa kích thước về 224×224 và chuyển đổi thành tensor để phù hợp với kiến trúc đầu vào của các mạng học sâu hiện đại. Chúng tôi sử dụng các kỹ thuật *tiền xử lý* và *tăng cường dữ liệu* (*data augmentation*) nhằm cải thiện khả năng khái quát hóa của mô hình đối với dữ liệu chưa từng thấy.

Cụ thể, các ảnh trong tập huấn luyện được áp dụng chuỗi các phép biến đổi ngẫu nhiên như sau:

- Lật ngang ngẫu nhiên với xác suất 0.5;
- Xoay ảnh trong khoảng $\pm 10°$;
- Tịnh tiến ngẫu nhiên theo cả hai trục trong phạm vi 10% kích thước ảnh;
- Thay đổi độ sáng, độ tương phản, độ bão hòa và sắc độ;
- Làm mờ Gaussian với kích thước kernel là 3 và $\sigma \in [0.1, 1.0]$;

Cuối cùng, tất cả ảnh được chuẩn hóa theo thông số ImageNet:

$$\text{mean} = [0.485,\ 0.456,\ 0.406], \quad \text{std} = [0.229,\ 0.224,\ 0.225].$$

Trong khi đó, các ảnh thuộc tập kiểm tra và kiểm định chỉ được thay đổi kích thước và chuẩn hóa, không áp dụng các kỹ thuật tăng cường dữ liệu nhằm đảm bảo tính khách quan trong đánh giá mô hình.

Tập dữ liệu được chia theo tỷ lệ xấp xỉ **8:1:1** cho ba tập **huấn luyện**, **kiểm tra** và **kiểm định**.

Tập dữ liệu	**FER2013**	**AffectNet**	**FER + AffectNet**
Train	31,709	17,793	31,671
Validation	3,589	2,202	3,379
Test	3,589	2,229	3,500
Tổng cộng	38,887	22,224	38,550

Bảng 4: So sánh số lượng ảnh theo từng tập giữa các bộ dữ liệu

4 Phương pháp

4.1 Động lực và lý do lựa chọn

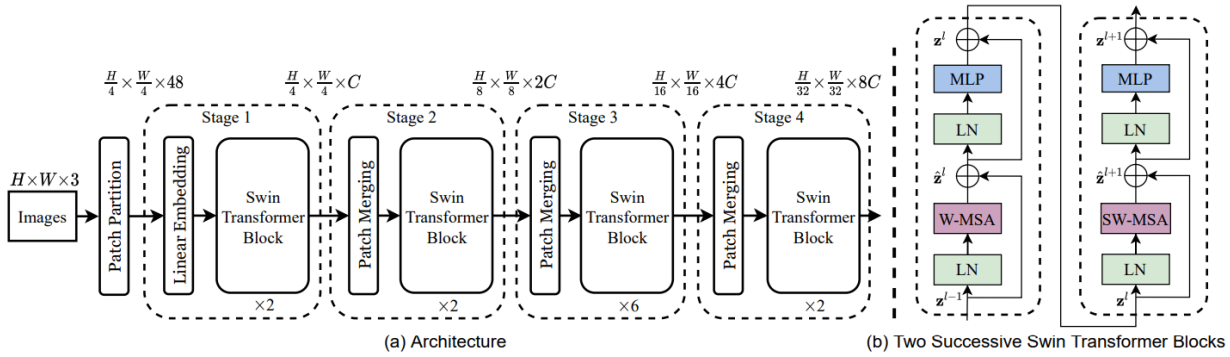
Gần đây, các phương pháp FER hiện đại chủ yếu dựa vào CNN như VGGNet, ResNet để trích xuất đặc trưng cục bộ. Tuy nhiên, CNN truyền thống bị giới hạn bởi receptive field cố định và thiếu khả năng khai thác quan hệ toàn cục. Trong khi đó, Transformer như Vision Transformer (ViT) cho thấy hiệu quả trong việc học quan hệ không gian toàn cục nhờ cơ chế self-attention, nhưng yêu cầu dữ liệu lớn và chưa tối ưu với đặc trưng cục bộ. Do đó, chúng tôi đề xuất kết hợp **Swin Transformer Tiny** và **ResNet-50**, nhằm tận dụng ưu điểm của cả hai mô hình. Ngoài ra, mô hình còn tích hợp SE block, Split Convolution và Mean Pooling để

tăng khả năng biểu diễn, giảm nhiễu và hạn chế overfitting.

4.2 Chi tiết các thành phần chính

4.2.1 Swin Transformer

Swin Transformer [3] là một kiến trúc Vision Transformer phân cấp, được thiết kế nhằm nâng cao hiệu quả tính toán trong xử lý ảnh. Khác với Vision Transformer (ViT) truyền thống, vốn sử dụng self-attention toàn cục dẫn đến độ phức tạp tính toán $\mathcal{O}(N^2)$ theo số lượng patch N, Swin Transformer thay thế bằng cơ chế attention theo cửa sổ cố định không chồng lắp (Window-based Multi-head Self-Attention - W-MSA), giúp giảm độ phức tạp xuống $\mathcal{O}(M^2 \cdot \frac{N}{M^2})$, trong đó M là kích thước cửa sổ.



Hình 3: (a) Kiến trúc Swin Transformer; (b) Hai khối liên tiếp với W-MSA và SW-MSA.

Để mô hình hóa hiệu quả mối quan hệ liên vùng, Swin Transformer giới thiệu kỹ thuật *Shifted Window* (SW-MSA), trong đó các cửa sổ được dịch chuyển giữa các tầng kế tiếp để đảm bảo khả năng truyền thông tin toàn cục mà vẫn giữ hiệu suất tính toán. Kiến trúc gồm bốn giai đoạn, kết hợp giữa patch embedding, patch merging và attention theo tầng, tạo ra biểu diễn phân cấp tương tự CNN.

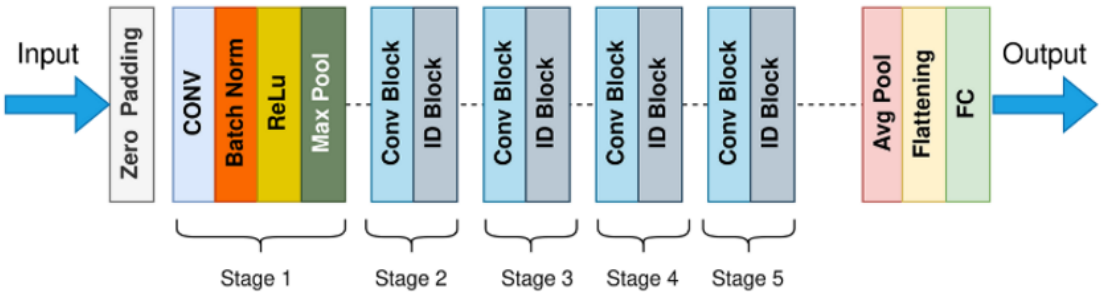
Mỗi khối Swin Transformer bao gồm: (1) attention theo cửa sổ (W-MSA hoặc SW-MSA), (2) một MLP hai tầng, (3) các lớp LayerNorm và (4) kết nối tắt

(residual connection), nhằm đảm bảo khả năng huấn luyện hiệu quả cho các mạng sâu.

4.2.2 ResNet-50

ResNet-50 [11] là một mạng nơ-ron tích chập sâu gồm 50 lớp, sử dụng cơ chế kết nối tắt (residual connection) để giảm thiểu hiện tượng mất mát gradient khi huấn luyện các mạng sâu. Kiến trúc bao gồm các khối residual với hai loại chính: identity block và convolutional block, cho phép mô hình học đặc trưng hiệu quả từ các chi tiết cục bộ như mắt, miệng và mũi – những vùng giàu thông tin trong bài toán nhận diện biểu cảm khuôn mặt.

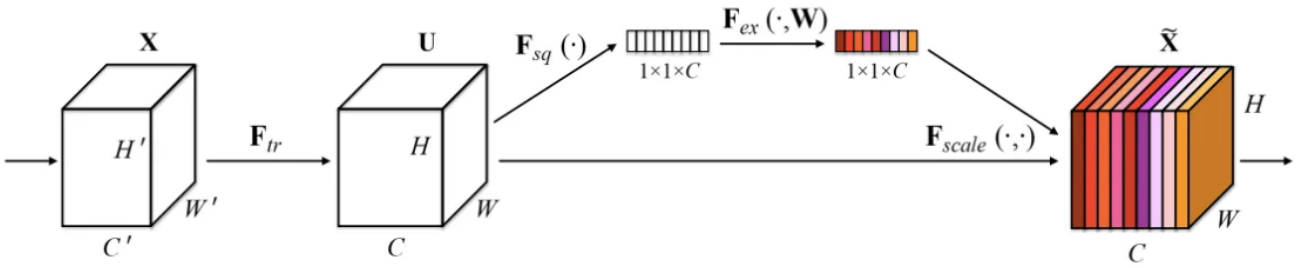
ResNet-50 đóng vai trò như một backbone trong nhánh đặc trưng cục bộ, hỗ trợ tăng độ ổn định trong huấn luyện và khả năng khái quát hoá cho mô hình.



Hình 4: Kiến trúc mạng ResNet-50.

4.2.3 Squeeze-and-Excitation (SE) Block

SE Block [6] là một mô-đun attention theo chiều kênh, nhằm tái phân phối trọng số các kênh đầu ra phù hợp với tầm quan trọng ngữ cảnh. Với hai giai đoạn chính – **Squeeze** và **Excitation**, SE block học được cách nhấn mạnh các kênh chứa thông tin phân biệt, đồng thời làm suy yếu các kênh không quan trọng. Đây là một kỹ thuật nhẹ, dễ tích hợp vào các mô hình CNN hiện có mà không làm tăng đáng kể độ phức tạp tính toán.



Hình 5: Cấu trúc SE Block.

Quá trình này bao gồm:

- **Squeeze:** Thực hiện *Global Average Pooling* trên toàn bộ không gian không gian để thu được vector đặc trưng $z \in \mathbb{R}^C$ – đại diện cho toàn bộ thông tin ngữ cảnh của mỗi kênh:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c}$$

  Điều này giúp mô hình tổng hợp thông tin không gian thành một biểu diễn toàn cục duy nhất cho từng kênh.

- **Excitation:** Vector z được đưa qua hai lớp fully connected liên tiếp để học cách gán trọng số cho từng kênh:

$$\mathbf{s} = \sigma(W_2 \cdot \delta(W_1 \cdot \mathbf{z}))$$

  Trong đó, δ là hàm kích hoạt ReLU, σ là hàm Sigmoid, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ và $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. Hệ số giảm r (thường là 16) giúp giảm chi phí tính toán và tránh overfitting.

Vector $\mathbf{s}$ chứa các hệ số quan trọng được sử dụng để hiệu chỉnh lại đầu vào theo từng kênh:

$$\hat{X}_c = s_c \cdot X_c$$

Việc nhân từng kênh với trọng số tương ứng giúp mô hình tập trung mạnh hơn vào các kênh có đóng góp phân biệt cao, đồng thời giảm tác động từ các kênh nhiễu không hữu ích.

Trong bối cảnh nhận diện cảm xúc khuôn mặt, SE Block đặc biệt hiệu quả vì cảm xúc thường được thể hiện rõ ràng qua một số khu vực và đặc trưng cụ thể trên khuôn mặt (ví dụ: mắt, miệng, chân mày).

4.2.4 Split Convolution và Mean Pooling

Để khai thác tốt hơn tính đa dạng của các kênh, mô hình sử dụng **Split Convolution** – chia tensor đầu vào $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ thành n nhóm theo chiều kênh:

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n], \quad \mathbf{X}_i \in \mathbb{R}^{H \times W \times \frac{C}{n}}$$

Mỗi phần được xử lý độc lập bằng một nhánh tích chập và kết quả được nối lại với nhau:

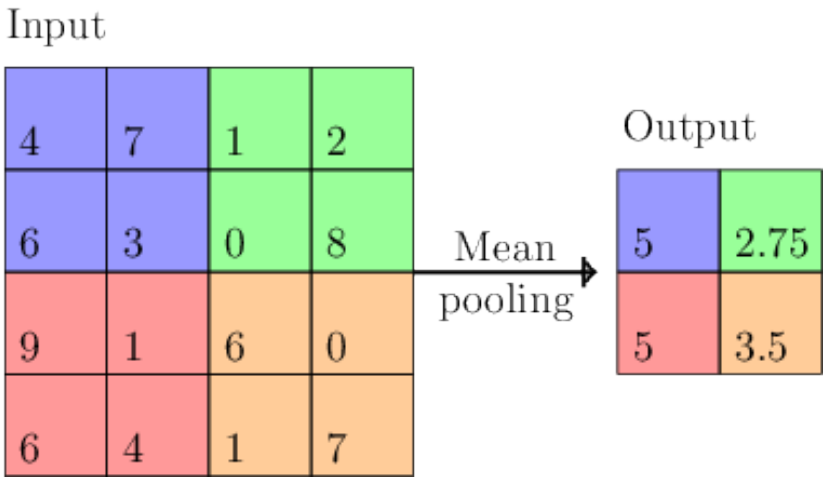
$$\mathbf{Y}_i = \text{Conv}_i(\mathbf{X}_i) \qquad ; \qquad \mathbf{Y} = \text{Concat}(\mathbf{Y}_1, ..., \mathbf{Y}_n)$$

Cách tiếp cận này có liên quan đến các kiến trúc như ResNeXt hoặc MobileNetv2, nơi việc chia nhỏ và xử lý kênh độc lập giúp tăng hiệu quả mô hình [11, 6].

Cuối cùng, **Mean Pooling** được áp dụng để giảm chiều và làm mượt không gian đặc trưng:

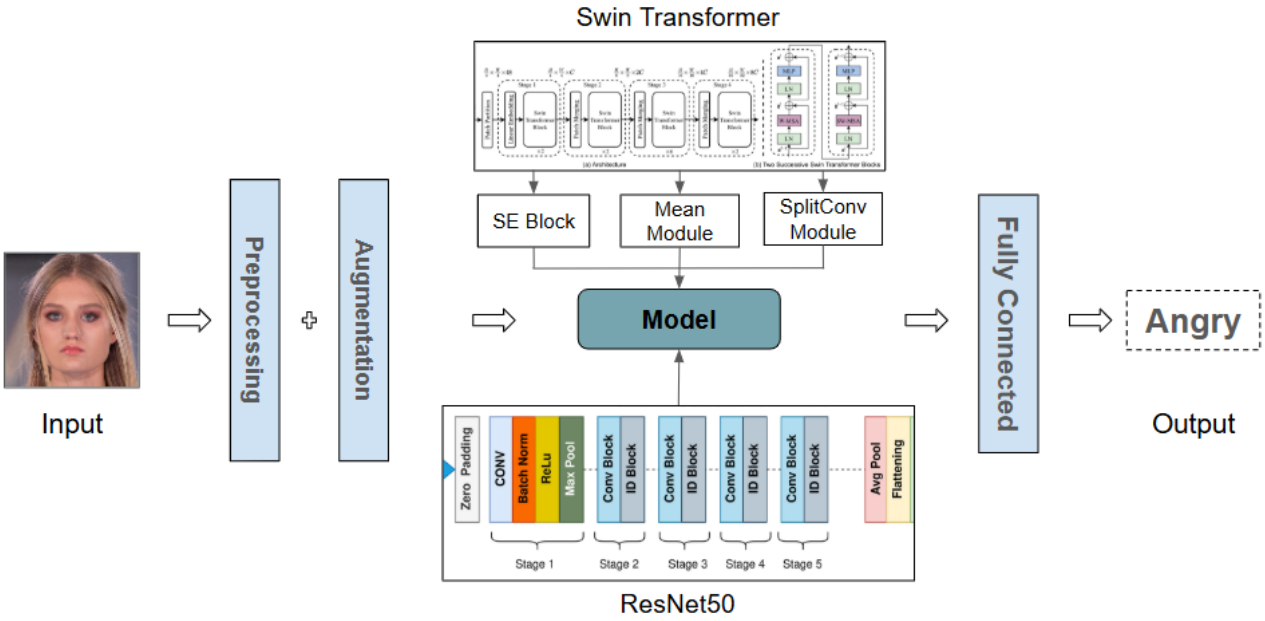
$$Y'_{i,j,c} = \frac{1}{k^2} \sum_{p=1}^{k} \sum_{q=1}^{k} Y_{i+p-1,j+q-1,c}$$

Tổ hợp này không những giảm số tham số mô hình mà còn giúp tăng tính ổn định, giảm overfitting trong giai đoạn huấn luyện [12].



Hình 6: Minh họa Mean Pooling.

4.3 Kiến trúc tổng thể



Hình 7: Pipeline mô hình Swin-SE-ResNet.

Mô hình Swin-SE-ResNet được thiết kế gồm hai nhánh chính nhằm khai thác cả đặc trưng toàn cục (global features) và đặc trưng cục bộ (local features) từ ảnh khuôn mặt:

- **Nhánh Swin Transformer:** Trích xuất đặc trưng toàn cục từ backbone Swin-Tiny pretrained. Đầu ra được đưa qua SE block, tiếp theo là Split Convolution và Mean Pooling để tăng khả năng biểu diễn và giảm nhiễu.
- **Nhánh ResNet-50:** Trích xuất đặc trưng cục bộ từ các vùng giàu thông tin của khuôn mặt. Đầu ra được xử lý qua Global Average Pooling để tạo vector biểu diễn.

Hai vector từ hai nhánh được nối lại và đưa qua lớp fully connected với hàm Softmax để dự đoán lớp cảm xúc:

$$\hat{\mathbf{y}} = \text{Softmax}(W \cdot \text{Concat}(\mathbf{f}_{\text{Swin}}, \mathbf{f}_{\text{ResNet}}) + b)$$

5 Thực nghiệm

5.1 Cấu hình huấn luyện

Mô hình được huấn luyện trên nền tảng Kaggle, sử dụng GPU NVIDIA Tesla P100 để tăng tốc quá trình tính toán và đảm bảo hiệu suất huấn luyện cho các mô hình học sâu với dữ liệu lớn.

Trong quá trình thực nghiệm, chúng tôi đã thử nghiệm nhiều bộ tham số khác nhau (batch size, learning rate và scheduler khác nhau) để đánh giá ảnh hưởng đến khả năng hội tụ và độ chính xác của mô hình. Cuối cùng, chúng tôi lựa chọn bộ tham số mang lại kết quả tốt nhất và ổn định nhất trong quá trình huấn luyện.

Dữ liệu được chia thành các mini-batch có kích thước batch size = 32, giúp tăng hiệu quả tính toán và đảm bảo sự ổn định trong cập nhật trọng số.

Mô hình sử dụng thuật toán tối ưu hóa Adam với learning rate khởi đầu là 1×10^{-4}. Learning rate này được điều chỉnh theo lịch trình StepLR với chu kỳ 10 epoch và hệ số giảm $\gamma = 0.1$, nhằm đảm bảo tốc độ học giảm dần đều, giúp mô hình hội tụ ổn định hơn về sau.

Tham số	Giá trị
Batch size	32
Optimizer	Adam
Learning rate	1×10^{-4}
StepLR	step size = 10, $\gamma = 0.1$
Loss function	Cross-Entropy Loss
Số epoch tối đa	50

Bảng 5: Cấu hình huấn luyện mô hình

Hàm mất mát được sử dụng là Cross-Entropy Loss, phù hợp với bài toán phân loại nhiều lớp. Mô hình được huấn luyện tối đa 50 epoch, với cơ chế early stopping nếu độ chính xác trên tập xác thực không được cải thiện sau 10 epoch liên tiếp. Điều này giúp tránh overfitting và tiết kiệm tài nguyên tính toán.

Sau mỗi epoch, mô hình được đánh giá trên tập xác thực theo ba chỉ số chính: Accuracy (độ chính xác), Loss (giá trị mất mát trung bình), và F1-score (trung bình có trọng số theo số lượng mẫu). Nếu độ chính xác trên tập xác thực đạt giá trị cao nhất từ trước đến nay, trọng số mô hình tại thời điểm đó sẽ được lưu lại.

Lịch sử huấn luyện được trực quan hóa bằng các biểu đồ thể hiện Accuracy, Loss và Validation F1-score theo từng epoch. Ngoài ra, một ma trận nhầm lẫn (confusion matrix) được xây dựng từ kết quả trên tập kiểm thử, giúp phân tích chi tiết hiệu suất của mô hình đối với từng lớp cảm xúc.

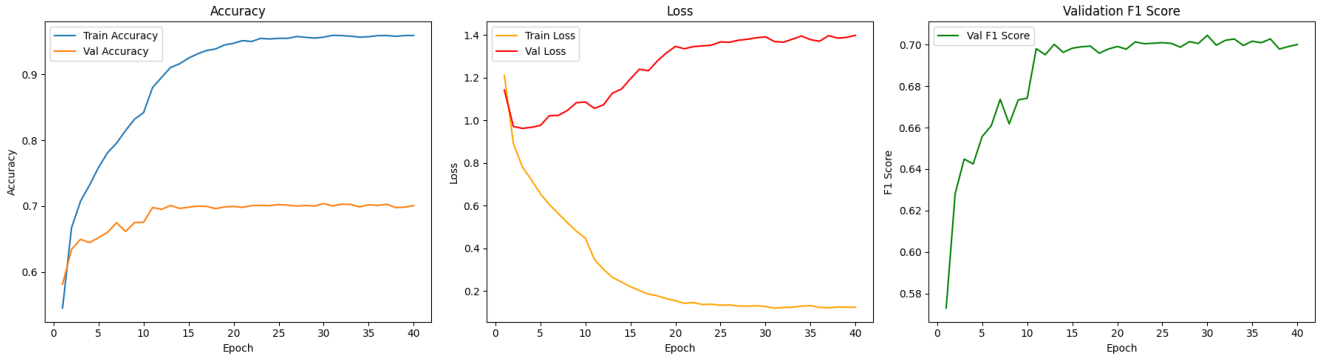
5.2 Kết quả thực nghiệm mô hình Swin-SE-ResNet

Kết quả huấn luyện trên tập FER2013

Quá trình huấn luyện dừng sớm tại epoch 40 do không ghi nhận sự cải thiện đáng kể nào trên tập validation trong vòng 10 epoch liên tiếp (theo tiêu chí Early Stopping). Tại epoch cuối cùng, mô hình đạt được:

- Loss huấn luyện: 0.1230 Độ chính xác: 0.9588
- Loss validation: 1.3988 Độ chính xác: 0.7008 F1-score: 0.7001

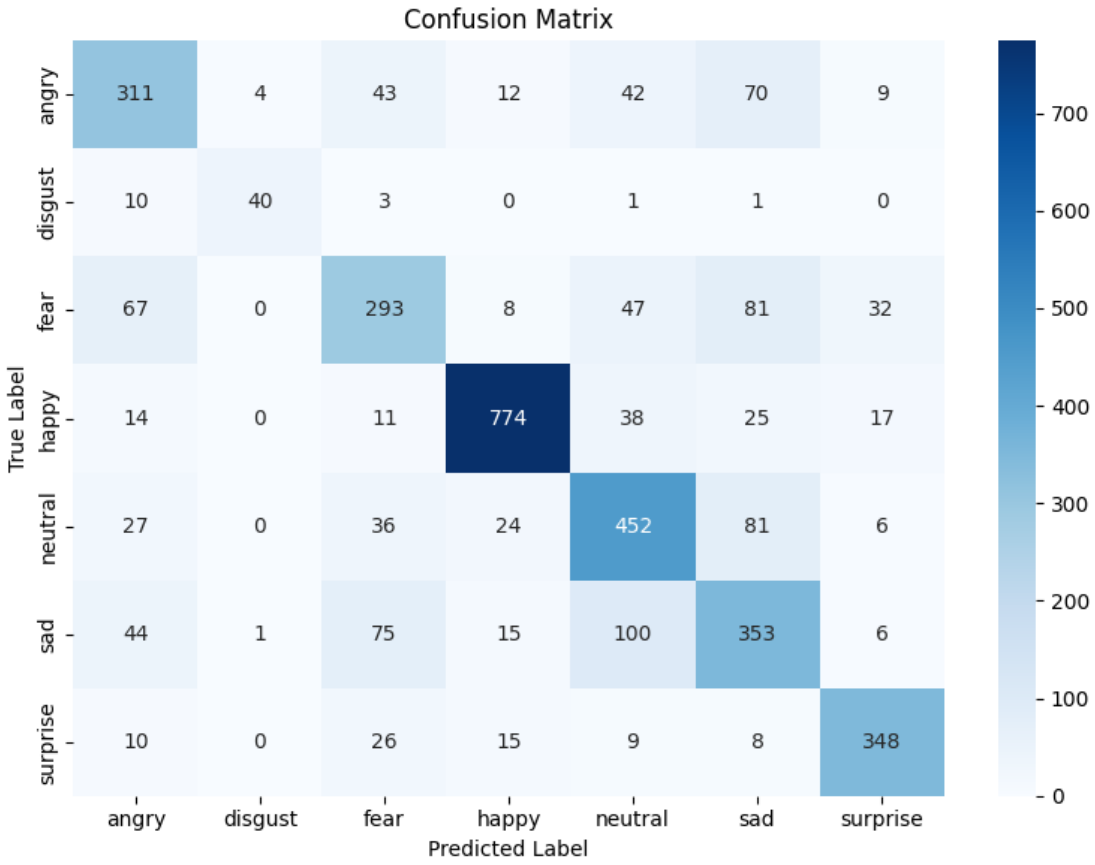
Ở hình 8, giai đoạn đầu huấn luyện (epoch 1–20) cho thấy hiệu suất mô hình được cải thiện rõ rệt, khi độ chính xác trên tập validation tăng từ 0.5815 lên 0.7013. Tuy nhiên, sau epoch 13, tốc độ cải thiện bắt đầu chậm lại. Dấu hiệu của hiện tượng quá khớp (overfitting) xuất hiện sau epoch 20, thể hiện qua việc loss trên tập validation bắt đầu tăng nhẹ trong khi loss trên tập huấn luyện vẫn tiếp tục giảm.



Hình 8: Quá trình huấn luyện trên FER2013

Sau quá trình đánh giá trên tập kiểm tra, mô hình đạt được kết quả **71.64%** và **71.71%** lần lượt tương ứng với độ chính xác và f1-score trung bình, thấp nhất trong ba tập dữ liệu. Đồng thời thực hiện đánh giá mô hình với kết quả ma trận nhầm lẫn như sau:

Từ hình 9, ta có thể thấy mô hình đạt hiệu suất nhận diện tốt nhất với cảm xúc **happy** (774 mẫu đúng), tiếp theo là **surprise** (348), và **neutral** (452). Điều này cho thấy mô hình có khả năng nhận diện các cảm xúc có biểu hiện rõ ràng và đặc trưng cao như vui vẻ và ngạc nhiên.



Hình 9: Ma trận nhầm lẫn trên tập kiểm tra FER2013

Tuy nhiên, mô hình vẫn gặp khó khăn trong việc phân biệt các cảm xúc tiêu cực có biểu cảm tương đối gần nhau. Ví dụ:

- **sad** thường bị nhầm lẫn với **fear** (75 mẫu), và **neutral** (100 mẫu).
- **angry** bị nhầm với **fear** (43 mẫu), **neutral** (42 mẫu), và **sad** (70 mẫu).
- **fear** cũng bị nhầm với nhiều lớp khác, đặc biệt là **sad** (81 mẫu) và **angry** (67 mẫu).

Một số lớp có số lượng mẫu thấp như **disgust** thường bị mô hình phân loại sai thành các lớp phổ biến hơn, chẳng hạn như **fear** hoặc **happy**. Điều này phản ánh sự mất cân bằng trong tập dữ liệu hoặc mô hình chưa học được đủ đặc trưng để phân biệt các cảm xúc ít xuất hiện.

Tóm lại, mô hình hoạt động tốt với các cảm xúc dễ nhận diện nhưng vẫn cần cải thiện khả năng phân biệt các cảm xúc tiêu cực tương đồng nhau.

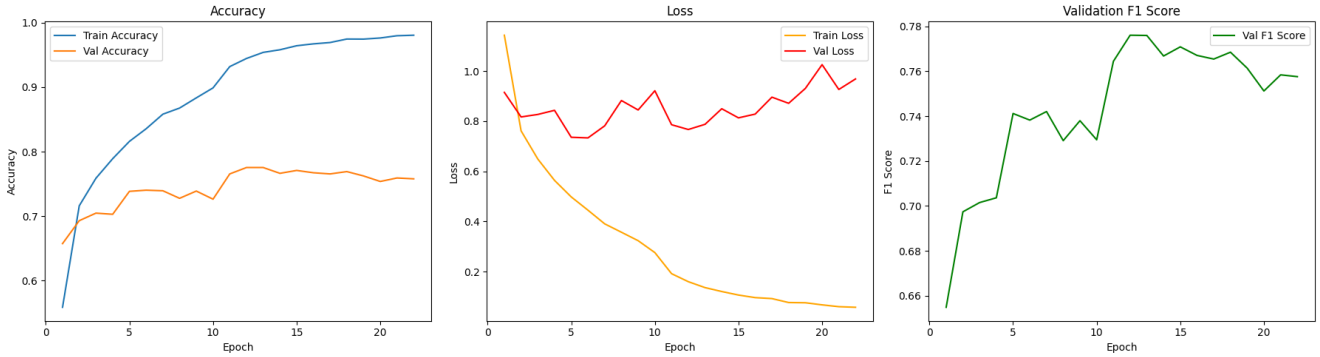
Để hiểu rõ hơn các trường hợp mô hình bị phân loại sai, chúng tôi trực quan hóa một số ví dụ ảnh từ tập kiểm tra FER2013 mà mô hình dự đoán sai nhãn.

Kết quả trên tập AffectNet

Huấn luyện mô hình trên tập AffectNet dừng lại ở epoch 22 do áp dụng cơ chế dừng sớm. Kết quả thu được như sau:

- Loss huấn luyện: 0.0578 Độ chính xác: 0.9804
- Loss validation: 0.9684 Độ chính xác: 0.7579 F1-score: 0.7576

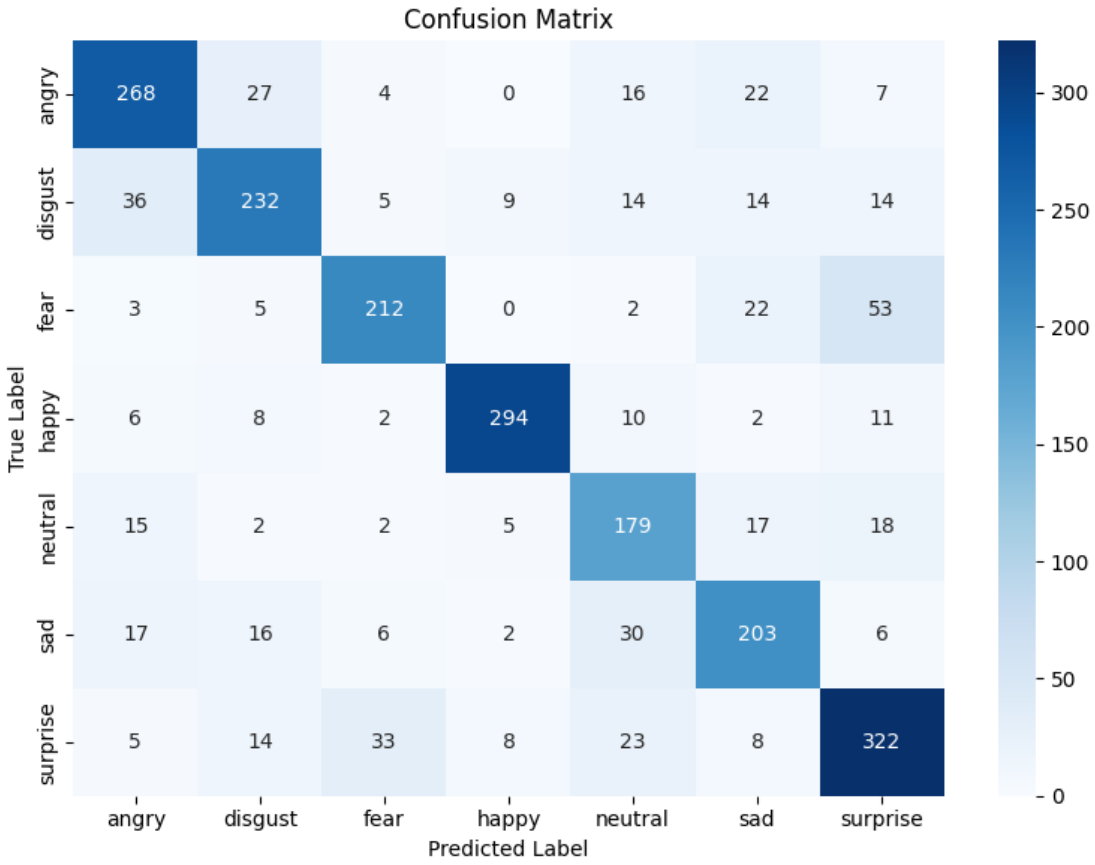
Mô hình học rất nhanh trong giai đoạn đầu (epoch 1–12), khi độ chính xác trên tập validation tăng mạnh từ 0.6575 lên 0.7754. Tuy nhiên, sau mốc này, tốc độ cải thiện giảm dần và xuất hiện hiện tượng overfitting nhẹ. Cơ chế Early Stopping được kích hoạt đúng thời điểm, giúp tránh mô hình bị học quá khớp với dữ liệu huấn luyện.



Hình 10: Quá trình huấn luyện trên AffectNet

Kết quả đánh giá trên tập AffectNet đạt kết quả cao nhất so với các tập dữ liệu khác, đạt được độ chính xác **76.72%** và f1-score trung bình **76.81%**.

Dựa vào hình 11, mô hình thể hiện hiệu suất nhận diện tốt nhất đối với lớp **happy** (294 mẫu), tiếp theo là **surprise** (322), **disgust** (232) và **angry** (268). Đây là những cảm xúc có biểu hiện khuôn mặt rõ ràng, giúp mô hình dễ học được các đặc trưng phân biệt.



Hình 11: Ma trận nhầm lẫn trên tập kiểm tra AffectNet

Tuy nhiên, các cảm xúc tiêu cực như **fear**, **sad**, và **neutral** lại bị nhầm lẫn tương đối nhiều:

- **Fear** thường bị nhầm với **surprise** (53 mẫu) và **sad** (22 mẫu), cho thấy mô hình khó phân biệt các biểu cảm có cường độ thấp và tương đồng về hình thái khuôn mặt.
- **Neutral** và **sad** dễ bị nhầm lẫn với nhau (30 mẫu), và đôi khi với **angry** hoặc **disgust**, phản ánh sự chồng lấn về biểu cảm giữa các trạng thái tiêu cực nhẹ.

- **Surprise** cũng bị nhầm lẫn một phần với **fear** (33 mẫu) và **happy** (23 mẫu), do các biểu cảm này đều có yếu tố mở rộng mắt, dễ gây nhiễu khi phân loại.

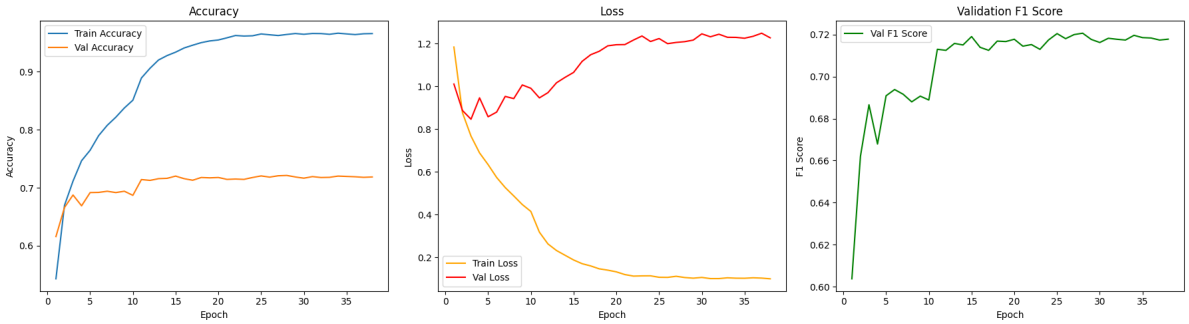
Nhìn chung, mô hình cho kết quả nhận diện khá tốt trên AffectNet, đặc biệt với các lớp cảm xúc có đặc trưng rõ ràng.

Kết quả trên tập kết hợp (FER2013 + AffectNet)

Trên tập dữ liệu kết hợp, mô hình được huấn luyện trong 38 epoch trước khi bị dừng sớm, với kết quả như sau:

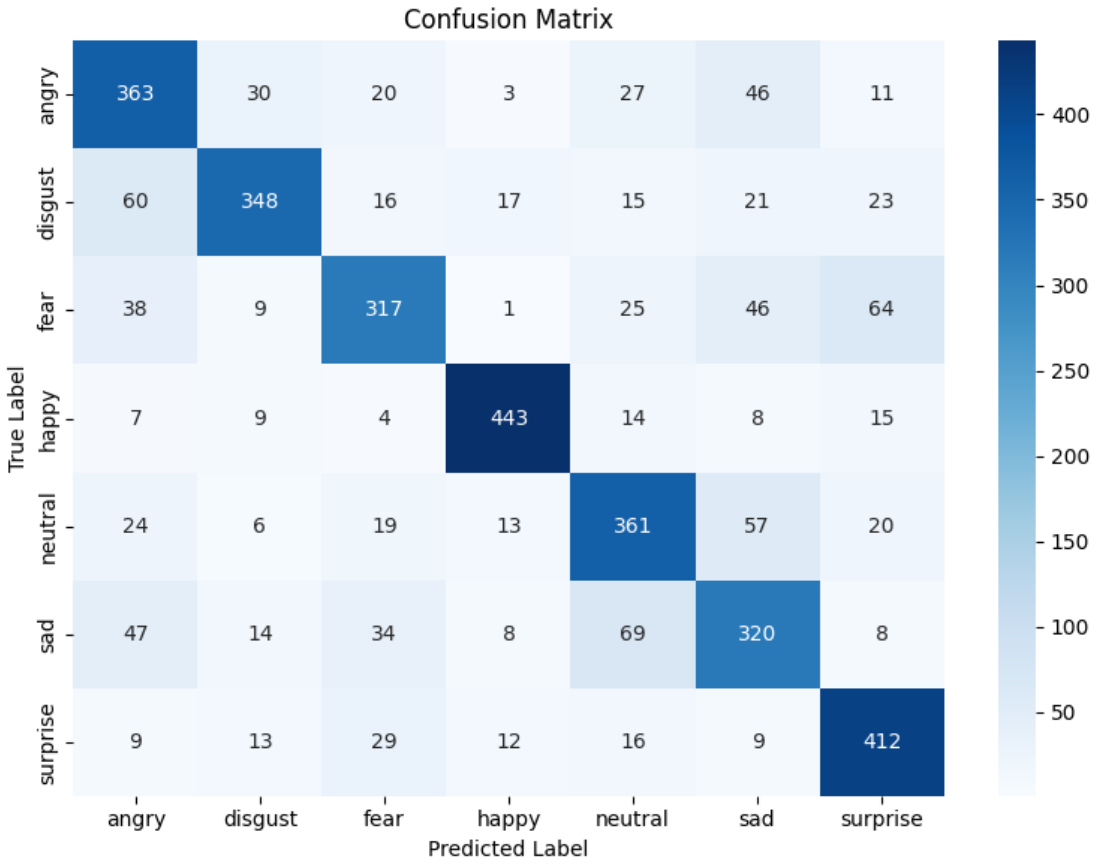
- Loss huấn luyện: 0.0995 Độ chính xác: 0.9656
- Loss validation: 1.2269 Độ chính xác: 0.7186 F1-score: 0.7178

Hiệu suất trên tập validation đạt mức tối ưu tại epoch 15 (accuracy = 0.7200, F1-score = 0.7190). Sau thời điểm này, mô hình tiếp tục tối ưu hóa trên tập huấn luyện nhưng bắt đầu xuất hiện overfitting nhẹ, thể hiện qua việc loss validation tăng dần.



Hình 12: Quá trình huấn luyện trên FER2013 + AffectNet

Kết quả đánh giá trên tập dữ liệu kết hợp đạt độ chính xác **73.26%** và F1-score trung bình **73.24%**. Đặc biệt, tập dữ liệu kiểm thử được xây dựng với phân bố đều giữa các lớp cảm xúc (500 ảnh mỗi nhãn), do đó kết quả phản ánh chính xác hơn khả năng tổng quát của mô hình trong các tình huống thực tế không bị thiên lệch dữ liệu.



Hình 13: Ma trận nhầm lẫn trên tập kiểm tra FER2013 + AffectNet

Hình 13 cho thấy mô hình huấn luyện trên tập dữ liệu kết hợp FER2013 và AffectNet đạt hiệu suất nhận diện tổng thể cao hơn so với từng tập riêng lẻ. Cụ thể:

- Các cảm xúc như **happy** (443 mẫu đúng), **surprise** (412), và **neutral** (361) tiếp tục là các lớp được phân loại chính xác cao nhất. Đây là bằng chứng cho thấy dữ liệu bổ sung từ AffectNet giúp cải thiện độ đa dạng và tính khái quát của mô hình.

- Mặc dù vẫn còn hiện tượng nhầm lẫn, đặc biệt giữa các cảm xúc tiêu cực như **fear** $\rightarrow$ **surprise** (64), hoặc **sad** $\leftrightarrow$ **neutral** (69), mức độ nhầm lẫn đã được giảm nhẹ so với khi huấn luyện trên AffectNet riêng lẻ.

- Đáng chú ý, lớp **fear** đạt 317 mẫu đúng – cao hơn đáng kể so với khi chỉ

dùng AffectNet (212), cho thấy lợi ích rõ rệt từ việc kết hợp thêm dữ liệu từ FER2013.

- Tuy nhiên, một số nhầm lẫn vẫn tồn tại như **angry** → **sad** (46) hoặc **disgust** → **angry** (60), phản ánh tính tương đồng biểu cảm giữa các cảm xúc tiêu cực mạnh.

Tổng thể, việc kết hợp hai tập dữ liệu giúp tăng cường tính đa dạng và tính khái quát cho mô hình, đồng thời cải thiện độ chính xác trên hầu hết các nhãn. Kết quả này khẳng định chiến lược mở rộng dữ liệu từ nhiều nguồn là một hướng tiếp cận hiệu quả trong bài toán nhận diện cảm xúc khuôn mặt.

Dưới đây là bảng tổng hợp kết quả thực nghiệm của mô hình trên ba tập dữ liệu:

Tập dữ liệu	**Accuracy**	**Weighted Avg. F1-Score**
FER2013	0.7164	0.7171
AffectNet	0.7672	0.7681
FER2013 + AffectNet	0.7326	0.7324

Bảng 6: Tổng hợp kết quả thực nghiệm trên các tập dữ liệu

5.3 So sánh với các mô hình khác

Hiệu suất của mô hình Swin-SE-ResNet được so sánh với các phương pháp hiện có trong lĩnh vực nhận diện cảm xúc khuôn mặt.

Kết quả trên tập dữ liệu FER2013 được trình bày trong bảng 7 cho thấy mô hình đề xuất đạt độ chính xác cao nhất so với các nghiên cứu trước đó.

Mô hình	Accuracy
CNN using the Adamax optimizer [16]	0.6600
VGG16+SEBlock [17]	0.6680
Swin-FER [5]	0.7111
Swin-SE-ResNet (Ours)	**0.7164**

Bảng 7: So sánh hiệu suất trên tập FER2013

Như thể hiện trong bảng 7, mô hình Swin-SE-ResNet đạt độ chính xác 71.64%, cao hơn 5.64% so với CNN sử dụng Adamax optimizer [16] và cao hơn 4.84% so với VGG16 kết hợp SE-block [17]. Đặc biệt, mô hình đề xuất còn vượt qua cả Swin-FER [5] – một phương pháp tối ưu hóa riêng cho bài toán nhận diện cảm xúc bằng Swin Transformer – với mức cải thiện 0.53%.

Những cải thiện này phản ánh hiệu quả tổng hợp của ba thành phần chủ đạo trong thiết kế mô hình: (1) năng lực trích xuất đặc trưng toàn cục của Swin Transformer thông qua cơ chế self-attention phân cấp, (2) khả năng tăng cường tập trung vào các đặc trưng kênh quan trọng nhờ khối SE, và (3) sự ổn định trong huấn luyện và khả năng truyền gradient sâu hiệu quả của kiến trúc ResNet.

Bên cạnh cấu trúc mô hình, kết quả tốt còn đến từ quy trình xử lý dữ liệu được tối ưu hóa: ảnh được chuẩn hóa về cùng kích thước và phân phối, kết hợp với chiến lược tăng cường dữ liệu đa dạng, bao gồm xoay, thay đổi độ sáng, độ tương phản và lật ảnh. Việc này không chỉ giúp mô hình làm quen với các biểu cảm ở nhiều điều kiện ánh sáng và góc chụp khác nhau, mà còn giảm đáng kể hiện tượng overfitting, từ đó cải thiện khả năng khái quát hóa trên dữ liệu chưa thấy.

Trên tập AffectNet, nhóm chúng tôi đã triển khai và huấn luyện các mô hình Swin Transformer (SwinT) và ResNet50 từ đầu như các baseline đối chứng, sử dụng cùng quy trình tiền xử lý và cấu hình huấn luyện. Kết quả thu được thể hiện trong bảng 8.

Mô hình	Accuracy	Weighted Avg. F1-Score
SwinT	0.7649	0.7645
ResNet50	0.7663	0.7666
SwinT-SE-SAM [15]	-	0.5420
TransFER [14]	0.6623	-
Swin-SE-ResNet (Ours)	**0.7672**	**0.7681**

Bảng 8: So sánh hiệu suất trên tập AffectNet

Mô hình Swin-SE-ResNet đạt accuracy 76.72% và F1-score 76.81%, cao nhất trong số các phương pháp được so sánh. Mặc dù mức cải thiện so với SwinT và ResNet50 do nhóm triển khai không lớn (chênh lệch dưới 1%), điều này vẫn phản ánh hiệu quả tích cực của việc tích hợp SE-block, giúp mô hình ưu tiên các đặc trưng kênh quan trọng hơn. Đồng thời, mô hình cũng cho thấy sự vượt trội rõ rệt so với các phương pháp phức tạp hơn như TransFER [14] và SwinT-SE-SAM [15], lần lượt kém hơn đến hơn 10% về accuracy và hơn 22% về F1-score.

Tổng quan, mô hình Swin-SE-ResNet không chỉ đạt hiệu suất cao mà còn duy trì độ ổn định trên các tập dữ liệu khác nhau. Kết quả này củng cố tính thực tiễn của kiến trúc đề xuất và nhấn mạnh vai trò quan trọng của việc kết hợp giữa mô hình mạnh và tiền xử lý tốt trong các bài toán thị giác máy tính.

5.4 Thử nghiệm mô hình real-time

Để đánh giá tính ứng dụng thực tiễn của mô hình, chúng tôi triển khai mô hình Swin-SE-ResNet trong một hệ thống nhận diện cảm xúc thời gian thực sử dụng webcam. Mục tiêu là kiểm tra khả năng phản hồi của mô hình trên dữ liệu ảnh động liên tục, đồng thời đánh giá tính ổn định khi áp dụng vào môi trường thật.

Cụ thể, hệ thống bao gồm các thành phần chính như sau:

- **Tiền xử lý ảnh**: Mỗi khung hình thu được từ webcam sẽ được lật ngược (mirror) và chuyển sang không gian màu RGB để xử lý. Bộ phát hiện khuôn mặt *MTCNN* [18] được sử dụng để phát hiện tất cả các khuôn mặt trong ảnh.

- **Tiền xử lý khuôn mặt**: Các vùng chứa khuôn mặt sau khi được trích xuất sẽ được resize về kích thước 224×224, chuyển thành tensor và chuẩn hóa theo thông số của ImageNet.

- **Mô hình nhận diện cảm xúc**: Mô hình được chọn là mô hình được huấn luyện trên tập dữ liệu **FER2013** — một tập dữ liệu lớn và đa dạng về cảm xúc khuôn mặt trong môi trường thực tế. Việc lựa chọn tập này giúp mô hình học được các đặc trưng phổ quát hơn và đạt hiệu quả tốt hơn khi triển khai thực tế.

- **Hiển thị kết quả**: Cảm xúc dự đoán được gán nhãn trên mỗi khuôn mặt trong khung hình, với *bounding box* màu xanh lá và nhãn màu xanh dương.

Quy trình xử lý được thực hiện liên tục trên từng khung hình, cho phép mô hình phản hồi theo thời gian thực. Trọng số của mô hình được tải từ tệp `.pth` đã được huấn luyện sẵn. Tuy nhiên, do mô hình có cấu trúc kết hợp giữa Swin Transformer, ResNet và SE-block, tốc độ xử lý trên CPU còn chậm, chưa đáp ứng tốt yêu cầu thời gian thực trong một số tình huống. Điều này mở ra nhu cầu tối ưu hóa mô hình hoặc chuyển sang các kiến trúc nhẹ hơn như MobileNet, EfficientNet hoặc triển khai trên GPU.

Kết quả thực nghiệm cho thấy mô hình có khả năng nhận diện tương đối chính xác các cảm xúc phổ biến và có đặc trưng rõ ràng như *happy*, *sad*, *angry* và *fear*. Đây là các cảm xúc có biểu hiện khuôn mặt nổi bật, dễ được trích xuất đặc trưng thông qua mạng lưới kết hợp Swin Transformer và SE-block. Tuy nhiên, các cảm xúc như *surprise*, *neutral* và *disgust* lại khó được phân biệt rõ ràng, do đặc điểm thị giác không đặc trưng hoặc dễ bị nhầm lẫn với các cảm xúc khác.

6 Kết luận

Trong đồ án này, chúng tôi đã đề xuất kiến trúc Swin-SE-ResNet cho bài toán nhận diện cảm xúc khuôn mặt và tiến hành đánh giá trên ba tập dữ liệu: FER2013, AffectNet và tập kết hợp FER2013+AffectNet. Mô hình đạt độ chính xác cao nhất trên cả ba tập so với các phương pháp tham khảo, với kết quả nổi bật nhất là accuracy 76.72% và F1-score 76.81% trên AffectNet, cùng với độ chính xác 73.26% trên tập dữ liệu kết hợp. Việc tích hợp Swin Transformer, SE-block và ResNet đã cho thấy hiệu quả tổng hợp rõ rệt, góp phần nâng cao hiệu suất nhận diện nhờ khả năng trích xuất đặc trưng mạnh mẽ, tập trung vào thông tin quan trọng và duy trì độ sâu huấn luyện ổn định. Đặc biệt, việc kết hợp dữ liệu từ FER2013 và AffectNet đã giúp cải thiện rõ rệt khả năng khái quát hoá của mô hình, giảm đáng kể hiện tượng nhầm lẫn ở các lớp cảm xúc tiêu cực như *fear*, *disgust* và *angry*.

Mặc dù mô hình đạt kết quả thấp hơn hai kiến trúc còn lại khi huấn luyện trên tập FER2013 đơn lẻ, nhưng trong quá trình triển khai ứng dụng thực tế thời gian thực với luồng dữ liệu liên tục và đa dạng, Swin-SE-ResNet lại thể hiện khả năng nhận diện ổn định và chính xác hơn. Điều này cho thấy tính linh hoạt và khả năng tổng quát hóa mạnh của kiến trúc, đặc biệt khi làm việc với dữ liệu ngoài phân phối (out-of-distribution) mà mô hình chưa từng được huấn luyện trực tiếp. Ngoài ra, khả năng thích nghi tốt với dữ liệu đầu vào phức tạp và điều kiện ánh sáng thay đổi cũng là một điểm mạnh đáng chú ý trong thực tế triển khai.

Hướng phát triển tương lai có thể bao gồm:

- Tối ưu hóa tốc độ suy luận của mô hình để triển khai hiệu quả hơn trên các thiết bị di động hoặc nhúng, phục vụ các ứng dụng yêu cầu phản hồi tức thời.
- Kết hợp thêm các nguồn dữ liệu đa dạng về văn hóa, giới tính và độ tuổi để nâng cao độ tin cậy và công bằng trong nhận diện cảm xúc trên phạm vi toàn cầu.
- Mở rộng phạm vi ứng dụng vào các lĩnh vực như hỗ trợ tâm lý, chăm sóc

sức khỏe tinh thần, giao tiếp người–máy (HCI), hoặc trong môi trường giáo dục thông minh nhằm theo dõi trạng thái cảm xúc của học sinh.

- Thử nghiệm với các kiến trúc nhẹ hơn như EfficientFormer, MobileViT hoặc EdgeNeXt nhằm cân bằng giữa độ chính xác và chi phí tính toán, đặc biệt phù hợp với hệ thống biên (edge computing).
- Tích hợp thêm các tín hiệu đa modal như giọng nói, chuyển động cơ thể hoặc văn bản để xây dựng hệ thống nhận diện cảm xúc toàn diện và chính xác hơn.

Tổng kết lại, mô hình Swin-SE-ResNet không chỉ mang lại kết quả tốt về mặt định lượng mà còn chứng minh tiềm năng ứng dụng rộng rãi trong các hệ thống tương tác thông minh theo thời gian thực, đặc biệt trong các kịch bản thực tế với dữ liệu phức tạp và liên tục.

Tài liệu

[1] Goodfellow, I., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. *International Conference on Neural Information Processing*, 117–124.

[2] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31.

[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

[4] Liang, X., Xu, L., Zhang, W., Zhang, Y., Liu, J., & Liu, Z. (2023). A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition. *The Visual Computer*, *39*, 2277–2290.

[5] Bie, M., Xu, H., Gao, Y., Song, K., & Che, X. (2024). Swin-FER: Swin Transformer for facial expression recognition. *Applied Sciences*, *14*(14), 6125.

[6] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132–7141.

[7] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6848–6856.

[8] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv preprint arXiv:1602.07360*.

[9] Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

[10] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929.

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

[12] Ma, F., Sun, B., & Li, S. (2021). Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, *14*(2), 1236–1248.

[13] Zhou, M., Liu, X., Yi, T., Bai, Z., & Zhang, P. (2023). A superior image inpainting scheme using Transformer-based self-supervised attention GAN model. *Expert Systems with Applications*, *233*, 120906.

[14] Xue, F., Wang, Q., & Guo, G. (2021). Transfer: Learning relation-aware facial expression representations with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3601–3610.

[15] Vats, A., & Chadha, A. (2023). Facial expression recognition using squeeze and excitation-powered Swin Transformers. *arXiv preprint arXiv:2301.10906v7*.

[16] Alamsyah, D., & Pratama, D. (2020). Implementasi Convolutional Neural Networks (CNN) untuk klasifikasi ekspresi citra wajah pada FER-2013 dataset. *Jurnal Teknologi Informasi*, *4*(2), 350–355.

[17] Nie, H. (2022). Face expression classification using Squeeze-Excitation based VGG16 network. *Proceedings of the 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Guangzhou, China, 14–16 January 2022, 482–485.

[18] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.