DOI: 10.64302/joshusc.v31n1e.1335

# SCoTER: MÔ HÌNH CNN-TRANSFORMER TINH GỌN NHẬN DẠNG CẢM XÚC KHUÔN MẶT

**Trần Thị Kiều[1*], Hồ Phước Tiến[2]**

[1] Khoa Điện, Điện tử và Công nghệ vật liệu, Trường Đại học Khoa học, Đại học Huế

[2] Khoa Điện tử-Viễn thông, Trường Đại học Bách Khoa, Đại học Đà Nẵng

*Email: kieutran@husc.edu.vn

*Ngày nhận bài: 8/9/2025; ngày hoàn thành phản biện: 02/10/2025; ngày duyệt đăng: 6/10/2025*

**TÓM TẮT**

Nhận dạng cảm xúc khuôn mặt (Facial Expression Recognition – FER) là một bài toán quan trọng trong thị giác máy tính, đóng vai trò then chốt trong các ứng dụng tương tác người–máy, giám sát hành vi và phân tích cảm xúc trong thời gian thực. Mặc dù các mô hình học sâu hiện nay đã đạt được nhiều tiến bộ, độ phức tạp tính toán cao vẫn là một thách thức lớn khi triển khai trong môi trường thực tế. Để giải quyết hạn chế này, nghiên cứu đề xuất một mô hình nhẹ (lightweight) được phát triển dựa trên ý tưởng của PAtt-Lite, nhằm nâng cao độ chính xác của mô hình đồng thời duy trì độ phức tạp tính toán ở mức hợp lý. Kết quả thực nghiệm cho thấy mô hình đạt 100% độ chính xác trên CK+ và 69% trên FER2013, cao hơn 7% so với mô hình gốc trên FER2013, điều này chứng minh tính hiệu quả và khả năng ứng dụng thực tế của mô hình được đề xuất.

**Từ khóa:** Nhận dạng cảm xúc khuôn mặt, MobileNetV1, Transformer, Patch Extraction, Mô hình nhẹ.

## 1. MỞ ĐẦU

Nhận dạng cảm xúc khuôn mặt (Facial Expression Recognition – FER) là một nhánh quan trọng của thị giác máy tính, tập trung vào việc trích xuất và phân tích trạng thái cảm xúc của con người từ hình ảnh hoặc video. Với các ứng dụng đa dạng trong tương tác người–máy, giám sát hành vi, giáo dục thông minh và y học tâm lý [1], [2], lĩnh vực này ngày càng thu hút sự quan tâm mạnh mẽ của cộng đồng nghiên cứu. FER không chỉ tập trung vào nhận dạng biểu cảm tĩnh mà còn hướng tới xử lý cảm xúc trong điều kiện môi trường tự nhiên. Điều này đặt ra yêu cầu về các mô hình phải có khả năng khái quát hóa tốt và duy trì độ ổn định cao khi triển khai trong thực tế.

Để đạt được khả năng khái quát hóa cao, các mô hình FER hiện nay vẫn phải đối mặt với nhiều thách thức quan trọng. Thứ nhất, trong môi trường thực tế, hệ thống nhận dạng cảm xúc dễ bị ảnh hưởng bởi các yếu tố như thay đổi tư thế, ánh sáng, mức độ che khuất và sự tương đồng giữa các biểu cảm. Thứ hai, các bộ dữ liệu phổ biến như FER2013 thường chứa nhiễu và mất cân bằng giữa các lớp cảm xúc, làm hạn chế khả năng học đặc trưng của mô hình. Thứ ba, những mô hình đạt độ chính xác cao thường đòi hỏi tài nguyên tính toán lớn, gây khó khăn cho việc triển khai trên thiết bị di động và trong các ứng dụng thời gian thực [3].

Để giải quyết những vấn đề trên, các mạng nơ-ron tích chập gọn nhẹ (lightweight CNN) như MobileNetV1 [4] và EfficientNet [5] đã được đề xuất, giúp giảm đáng kể số lượng tham số mà vẫn duy trì hiệu quả trích xuất đặc trưng. Các mô hình này hướng đến việc cân bằng giữa độ chính xác và hiệu suất tính toán. Trong nghiên cứu [6], tác giả phát triển CLCM và đánh giá trên FER2013, đạt 63%, so với ShuffleNetV2 (65%) và MobileNetV2 (58%). Bên cạnh các mô hình gọn nhẹ, nhiều nghiên cứu cũng thử nghiệm những mô hình có số tham số lớn nhằm cải thiện hiệu suất nhận dạng, nhưng cơ chế Self-Attention đơn. Kết quả này cho thấy việc tăng độ sâu hoặc mở rộng số lượng tham số không đảm bảo cải thiện hiệu quả trên dữ liệu FER "in-the-wild", do hiện tượng nhiễu, mất cân bằng lớp và sai nhãn vẫn ảnh hưởng mạnh đến quá trình học đặc trưng.

Để khắc phục những hạn chế của FER, nhiều nghiên cứu gần đây đã tập trung phát triển các mô hình gọn nhẹ kết hợp cơ chế chú ý, nhằm đạt được sự cân bằng giữa hiệu suất tính toán và khả năng khái quát hóa. Cụ thể, các mạng nơ-ron tích chập gọn nhẹ kết hợp với khối trích xuất vùng (Patch Extraction Block) giúp giảm đáng kể số lượng tham số và tăng khả năng nắm bắt đặc trưng cục bộ, trong khi khối Transformer với cơ chế Multi-Head Attention (MHA) nhằm nâng cao hiệu quả nhận dạng cảm xúc trong môi trường phức tạp và "in-the-wild" [8].Một trong những mô hình tiêu biểu được phát triển theo hướng tiếp cận trên là PAtt-Lite [3], được thiết kế nhằm giải quyết các hạn chế về độ phức tạp và khả năng trích xuất đặc trưng trong FER. Tuy nhiên, do chỉ sử dụng cơ chế Self-Attention đơn giản, khả năng khái quát hóa của mô hình còn hạn chế, đặc biệt khi áp dụng trên các tập dữ liệu "in-the-wild" như FER2013. Để cải thiện, nghiên cứu này đề xuất SCoTER (Small Convolutional-Transformer for Emotion Recognition), thay Self-Attention bằng Transformer với MHA, điều chỉnh vị trí Global Average Pooling (GAP) và thiết kế lại Patch Extraction Block để học song song nhiều đặc trưng cục bộ.

Kết quả thực nghiệm cho thấy cả mô hình PAtt-Lite và mô hình đề xuất SCoTER đạt 100% độ chính xác trên tập dữ liệu CK+ (https://www.kaggle.com/datasets/davilsena/ckdataset) phản ánh khả năng nhận dạng biểu cảm tốt trong môi trường có kiểm soát, trên tập FER2013, mô hình đề xuất đạt 69%

độ chính xác, cao hơn 7% so với PAtt-Lite (62%). Sự cải thiện này chứng tỏ việc tích hợp khối Transformer sử dụng cơ chế Multi-Head Attention (MHA) cùng với thiết kế lại vị trí lớp GAP và khối Patch Extraction đã giúp mô hình nâng cao khả năng khái quát hóa và hiệu suất nhận dạng cảm xúc, đồng thời duy trì độ phức tạp tính toán ở mức hợp lý cho các ứng dụng thời gian thực và thiết bị giới hạn phần cứng.

Trong phần tiếp theo của bài báo, mục 2 giới thiệu những kiến thức cơ bản về CNN và Transformer. Mục 3 sẽ mô tả chi tiết phương pháp đề xuất. Kết quả thực nghiệm và nhận xét sẽ được trình bày trong mục 4. Cuối cùng, mục 5 tóm tắt những kết quả đạt được của mô hình đề xuất và gợi ý hướng phát triển tiếp theo.

## 2. CƠ SỞ LÝ THUYẾT

### 2.1. Mạng CNN

Mạng nơ-ron tích chập (CNN) là một mô hình then chốt trong lĩnh vực thị giác máy tính, mô phỏng cơ chế xử lý hình ảnh của con người thông qua các lớp tích chập (convolutional layers) [9], [10]. Mạng có khả năng tự động học các đặc trưng ở nhiều mức độ — từ biên, góc, kết cấu đến hình dạng và biểu cảm khuôn mặt. Đặc biệt, trong bài toán nhận dạng cảm xúc khuôn mặt (FER), CNN giúp mô hình tập trung vào các vùng đặc trưng giàu thông tin như mắt, miệng và lông mày, giảm sự phụ thuộc vào các phương pháp trích chọn thủ công, đồng thời nâng cao khả năng khái quát hóa trước các biến thiên về tư thế và ánh sáng. Tuy nhiên, các mô hình CNN sâu như VGGNet [9] và ResNet [10] có số lượng tham số lớn, gây khó khăn khi triển khai trong môi trường giới hạn phần cứng. Để khắc phục, MobileNetV1 [4] áp dụng kỹ thuật *depthwise separable convolution*. Cách tiếp cận này giúp giảm đáng kể chi phí tính toán trong khi vẫn duy trì hiệu quả trích xuất đặc trưng, phù hợp cho các mô hình FER gọn nhẹ và hiệu quả.

### 2.2. Vision transformers

Mặc dù các mạng CNN gọn nhẹ như MobileNetV1 [4] đã chứng minh hiệu quả trong việc trích xuất đặc trưng cục bộ với chi phí tính toán thấp, chúng vẫn gặp hạn chế trong việc nắm bắt mối quan hệ toàn cục giữa các vùng trên ảnh, do bản chất phép tích chập chỉ khai thác thông tin trong phạm vi cục bộ. Để khắc phục hạn chế này, các nghiên cứu gần đây đã hướng tới việc kết hợp cơ chế *Attention* - tiêu biểu là Vision Transformer (ViT) [8] - nhằm tăng khả năng biểu diễn không gian toàn cục của mô hình. Tuy nhiên, ViT thường yêu cầu lượng dữ liệu huấn luyện lớn và tài nguyên tính toán đáng kể, khiến việc áp dụng trong các tác vụ quy mô nhỏ như nhận dạng cảm xúc khuôn mặt (FER) trở nên khó khăn. Vì vậy, các mô hình gần đây — điển hình là PAtt-Lite [3] — đã đề xuất kết hợp kiến trúc CNN gọn nhẹ và cơ chế Attention, tận dụng đồng thời ưu điểm của cả hai: khả năng học đặc trưng cục bộ mạnh mẽ của CNN và khả năng biểu diễn toàn cục

của Transformer, giúp cải thiện hiệu suất nhận dạng trong khi vẫn duy trì mức độ tính toán thấp.

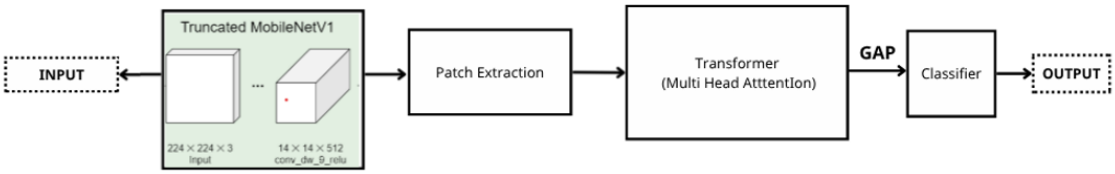
## 3. PHƯƠNG PHÁP ĐỀ XUẤT

### 3.1 Kiến trúc mô hình

Dựa trên cơ sở của các mạng CNN gọn nhẹ và cơ chế Attention trong Vision Transformer, nhóm nghiên cứu đề xuất kiến trúc mô hình nhận dạng cảm xúc SCoTER (Hình 1), kế thừa ý tưởng từ PAtt-Lite [3] nhưng được điều chỉnh và tối ưu hóa nhằm tăng hiệu quả biểu diễn trong khi vẫn đảm bảo tính gọn nhẹ.

Kiến trúc SCoTER bao gồm một phiên bản MobileNetV1 rút gọn (Truncated MobileNetV1) được sử dụng làm backbone của mô hình. Bản đồ đặc trưng thu được từ Truncated MobileNetV1 sau đó được đưa vào khối trích xuất vùng (Patch Extraction Block), nơi chúng được chia thành các patch đặc trưng cố định, đa dạng cho các vùng biểu cảm khác nhau trên khuôn mặt. Các patch này tiếp tục được xử lý bởi khối Transformer, giúp mô hình hóa mối quan hệ toàn cục giữa các vùng biểu cảm thông qua cơ chế Multi-Head Attention.

Tổng thể, mô hình bao gồm năm thành phần chính: 1) Truncated MobileNetV1 – trích xuất đặc trưng ban đầu, 2) Patch Extraction – chia bản đồ đặc trưng thành các vùng nhỏ, 3) Transformer – mô hình hóa quan hệ toàn cục, 4) Global Average Pooling (GAP) – nén đặc trưng toàn cục, 5) Classifier – phân loại cảm xúc đầu ra.



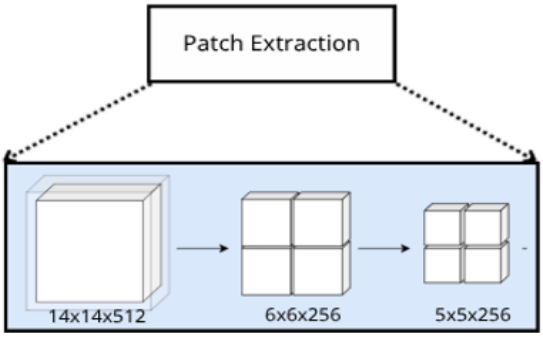

***Hình 1.*** Kiến trúc mô hình SCoTER

#### 3.1.1 MobileNetV1

Trong mô hình SCoTER, nhóm tác giả sử dụng phiên bản rút gọn của MobileNetV1 (Truncated MobileNetV1) làm mạng nền (backbone) cho khối trích xuất đặc trưng, được minh họa trong **Hình 1**. Cấu trúc này giúp giảm đáng kể độ phức tạp tính toán nhưng vẫn duy trì khả năng học đặc trưng cục bộ cho các vùng biểu cảm trên khuôn mặt. Đầu ra của mạng Truncated MobileNetV1 là bản đồ đặc trưng kích thước 14 × 14 × 512, chứa các thông tin biểu cảm quan trọng phục vụ cho các khối xử lý tiếp theo, đặc biệt là khối chia vùng đặc trưng (Patch Extraction) và khối Transformer.

### 3.1.2 Patch Extraction Block

Dựa trên cấu trúc của PAtt-Lite [3] và ý tưởng từ Vision Transformer (ViT) [8], nghiên cứu đề xuất Patch Extraction Block được hiệu chỉnh nhằm nâng cao khả năng học đặc trưng và chuyển đổi thông tin không gian sang dạng tuần tự phù hợp cho Transformer.



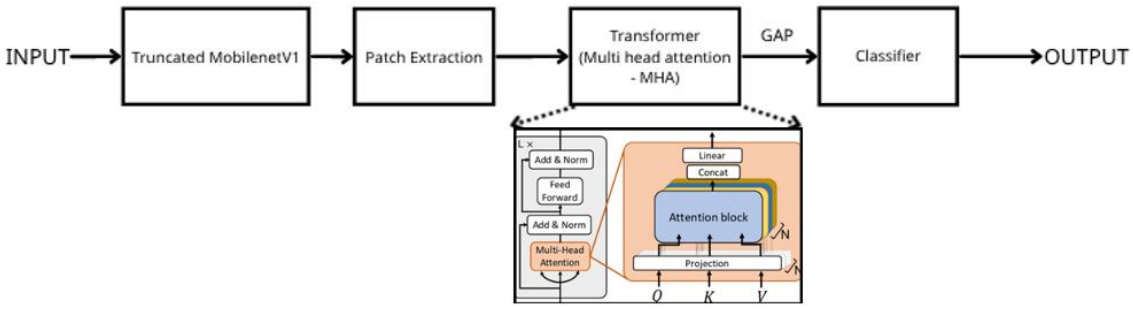

***Hình 2.*** Cấu trúc chi tiết của khối Patch Extraction

Đầu ra của MobileNetV1 (14×14×512) được đưa trực tiếp vào khối Patch Extraction Block, được thiết kế gồm ba lớp tích chập liên tiếp (Hình 2): hai lớp đầu là depthwise separable convolution, và lớp cuối là pointwise convolution (1×1).

Thiết kế này giúp mô hình duy trì khả năng học đặc trưng không gian hiệu quả của CNN, đồng thời chuyển đổi đặc trưng sang dạng tuần tự tương thích với Transformer, tạo liên kết hiệu quả giữa MobileNetV1 và khối Transformer.

### 3.1.3 Transformer Block

Tiếp nối quá trình trích xuất đặc trưng bằng MobileNetV1 và tái tổ chức không gian đặc trưng thông qua Patch Extraction Block, các patch đầu ra được đưa vào Transformer Encoder Block – thành phần được đề xuất nhằm tăng cường khả năng học biểu cảm toàn cục. Khác với PAtt-Lite [3], nghiên cứu này áp dụng Transformer Encoder Block hoàn chỉnh [11] (Hình 3). Nhờ cải tiến này, mô hình không chỉ duy trì khả năng học thông tin cục bộ từ từng patch, mà còn có thể mô hình hóa các mối quan hệ toàn cục giữa các vùng biểu cảm – điều mà cơ chế Self-Attention đơn trong PAtt-Lite [3] chưa thể hiện đầy đủ. Trong nghiên cứu này chúng tôi sử dụng MHA với 8 heads cho phép mô hình học đồng thời nhiều kiểu quan hệ không gian khác nhau giữa các patch. Qua đó, mô hình đề xuất tận dụng hiệu quả nền tảng gọn nhẹ của MobileNetV1, đồng thời nâng cao đáng kể hiệu suất nhận dạng cảm xúc trên các tập dữ liệu FER.



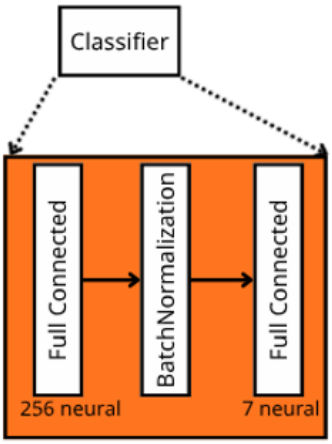

***Hình 3.*** Cấu trúc chi tiết của khối Transformer [11]

### 3.1.4 Global Average Pooling (GAP)

Sau khi các đặc trưng biểu cảm được học và tái cấu trúc thông qua Transformer Encoder Block, đầu ra của khối này được đưa vào Global Average Pooling (GAP) [10] để tổng hợp thông tin toàn cục. Trong PAtt-Lite [3], GAP được đặt ngay sau Patch Extraction Block, tức là quá trình nén đặc trưng diễn ra trước khi mô hình học mối quan hệ giữa các patch. Cách sắp xếp này giúp duy trì mô hình gọn nhẹ, nhưng dẫn đến mất mát thông tin không gian giữa các vùng biểu cảm - yếu tố quan trọng đối với bài toán nhận dạng cảm xúc. Trong nghiên cứu này, GAP được giữ nguyên về chức năng nhưng điều chỉnh vị trí (Hình 1), được đặt sau Transformer Encoder Block. Nhờ đó, các đặc trưng đầu vào đã được mô hình hóa đầy đủ mối quan hệ toàn cục trước khi thực hiện phép tổng hợp trung bình, giúp tăng hiệu quả biểu diễn mà vẫn duy trì độ phức tạp tính toán thấp.

### 3.1.5 Classifier Block

Sau khi các đặc trưng toàn cục được tổng hợp bằng Global Average Pooling (GAP), đầu ra được chuyển vào Classifier Block để dự đoán nhãn cảm xúc. Trong mô hình được đề xuất, khối **Classifier** Block được thiết kế gồm ba lớp liên tiếp để thực hiện việc phân loại sau khi các đặc trưng được trích xuất. Lớp đầu tiên là một lớp Dense với 128 hoặc 256 nơ-ron, chịu trách nhiệm học các ánh xạ phức tạp từ đầu ra của lớp Global Average Pooling (GAP). Tiếp theo là lớp BN - Batch Normalization giúp chuẩn hóa phân bố đặc trưng, ổn định quá trình huấn luyện và hạn chế overfitting. Cuối cùng một lớp Dense với 7 nơ-ron đầu ra, tương ứng với số lượng lớp cảm xúc, kết hợp với hàm Softmax để sinh xác suất dự đoán cuối cùng. Nhờ đó, mô hình hội tụ nhanh hơn và cải thiện độ chính xác nhận dạng cảm xúc trên các tập dữ liệu FER2013 và CK+.




***Hình 4.*** Cấu trúc chi tiết của khối Classifier

## 4. KẾT QUẢ THỰC NGHIỆM

### 4.1. Dữ liệu

Nghiên cứu này sử dụng hai bộ dữ liệu chuẩn CK+ và FER2013 nhằm đánh giá mô hình trong cả môi trường có kiểm soát và điều kiện tự nhiên. Cách lựa chọn này giúp phản ánh đồng thời độ chính xác nhận dạng và khả năng khái quát hóa của mô hình.

CK+ [12] là bộ dữ liệu thu thập trong môi trường phòng thí nghiệm có kiểm soát, gồm 593 chuỗi hình ảnh từ 123 đối tượng, trong đó 327 chuỗi hình được gán nhãn theo bảy cảm xúc cơ bản: **Anger, Disgust, Fear, Happiness, Sadness, Surprise và Contempt**. Trong bài báo này, nhóm tác giả sử dụng bộ dữ liệu bao gồm 981 hình ảnh (posed-dataset), được lấy từ 327 chuỗi hình ảnh của tập CK+ [12], chia thành 784 ảnh huấn luyện, 99 ảnh kiểm thử và 98 ảnh kiểm định. Dữ liệu có chất lượng hình ảnh cao, ánh sáng và tư thế được chuẩn hóa, thường được sử dụng làm chuẩn đánh giá trong các điều kiện lý tưởng.

FER2013 [13] bao gồm 35.887 ảnh khuôn mặt xám (48×48 pixel) được thu thập tự động từ Internet, trong đó, gồm 28709 ảnh dùng cho huấn luyện, 3589 ảnh cho kiểm thử và 3589 ảnh cho kiểm định Mỗi ảnh được gán nhãn theo bảy loại cảm xúc cơ bản, tuy nhiên do chỉ được gán nhãn bởi một người đánh giá, bộ dữ liệu có thể chứa nhiễu nhãn, phản ánh rõ tính đa dạng và phức tạp của dữ liệu thực tế.

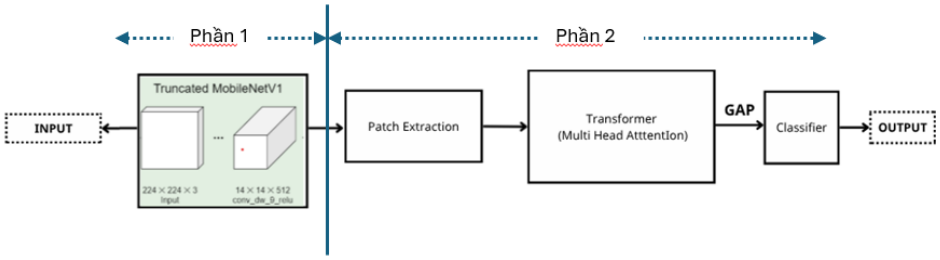
### 4.2. Huấn luyện mô hình

Tương tự như cách thức huấn luyện ở [3], quá trình huấn luyện mô hình đề xuất được chia thành **hai giai đoạn chính** thực hiện trên hai phần (Phần 1 và Phần 2) của mô hình được thể hiện ở Hình 5.

- Giai đoạn 1 – Huấn luyện phần 2: Giữ cố định (đóng băng) phần Truncated MobileNetV1 đã được huấn luyện sẵn trên ImageNet, chỉ huấn luyện các khối

phía sau gồm *Extraction Patch, Transformer (Multi Head Attention – MHA), GAP* và *Classifier*.

- Giai đoạn 2 – Tinh chỉnh mô hình cả Phần 2 và Phần 1: Mở khóa (unfreeze) các lớp trong MobileNetV1(Phần 1) và huấn luyện toàn bộ mô hình với tốc độ học nhỏ hơn nhằm điều chỉnh nhẹ các trọng số đã học mà không làm mất ổn định bộ tham số ban đầu.




***Hình 5.*** Cấu trúc chi tiết của Phần 1 và Phần 2 của mô hình SCoTER

Chi tiết của hai giai đoạn huấn luyện mô hình được thể hiện ở **Bảng 1.**

***Bảng 1.*** Chi tiết quá trình huấn luyện của giai đoạn 1 và giai đoạn 2

| Thành phần | Giai đoạn 1 | Giai đoạn 2 |
|---|---|---|
| **Mục tiêu** | Thích nghi đặc trưng MobileNetV1 với dữ liệu cảm xúc. | Đồng bộ hóa trọng số giữa backbone và phần nhận dạng cảm xúc |
| **Dữ liệu** | CK+ (kiểm soát) & FER-2013 (tự nhiên) | CK+ (kiểm soát) & FER-2013 (tự nhiên) |
| **Phần 1** | Khoá toàn bộ mô hình Truncated MobileNetV1 | Mở khóa một phần Truncated MobileNetV1: CK+ từ lớp 40, FER-2013 từ lớp 46 |
| **Phần 2** | Extraction Patch → Transformer (1–2 block, 4 heads, FF-dim 256/512) → GAP → Classifier (Dense 256, Dropout 0.1) | Extraction Patch → Transformer (1–2 block, 4 heads, FF-dim 256/512) → GAP → Classifier (Dense 256, Dropout 0.3) |
| **Bộ tham số** | • LR: 0.01, giảm 0.5× nếu 8 epoch không cải thiện >0.003, sau 20 epoch không cải thiện thì sẽ dừng quá trình huấn luyện.<br>• Thuật toán: Adam<br>• Loss: Categorical Cross-Entropy<br>• Max epoch: 200 | • LR: 0.001, giảm 0.5× nếu 10 epoch không cải thiện >0.002, sau 20 epoch không cải thiện thì sẽ dừng quá trình huấn luyện.<br>• Thuật toán: Adam<br>• Loss: Categorical Cross-Entropy<br>• Max epoch: 500 |

### 4.3. Kết quả và nhận xét

Kết quả đánh giá của mô hình đề xuất trên tập kiểm tra của bộ dữ liệu CK+ và FER2013 được thể hiện ở **Bảng 2**. Ở đây, mô hình đề xuất sử dụng 1 block Transformer, FF-dim = 256. Kết quả cho thấy, cũng tương tự như mô hình PAtt-Lite [3], mô hình đề xuất cho kết quả nhận dạng đúng là 100% với tập dữ liệu CK+. Với tập dữ liệu FER2013, phức tạp và đa dạng hơn, mô hình đề xuất đã thể hiện được ưu điểm so với mô hình PAtt-Lite [3], cũng như những mô hình khác như MobileNetV2 [6] và CLCM [6].

***Bảng* 2.** So sánh độ chính xác với các mô hình tiên tiến khác trên bộ FER2013 và CK+

| Mô hình | Số tham số | Bộ dữ liệu | Độ chính xác trên tập đánh giá |
|---|---|---|---|
| PAtt-Lite | 1,1M | CK+ | 100% |
| SCoTER (Model 1) | 1,5M | CK+ | 100% |
| SCoTER (Model 1) | 1,5M | FER2013 | 67% |
| PAtt-Lite | 1,1M | FER2013 | 62% |
| MobileNet V2 [6] | 2,4M | FER2013 | 58% |
| CLCM [6] | 3.9M | FER2013 | 65% |

Bên cạnh đó, để khảo sát sự ảnh hưởng của số lượng khối Transformer và số chiều FF-dim trong Transformer, **Bảng 3** so sánh độ chính xác trên tập FER2013 của 4 biến thể của mô hình đề xuất gồm: 1 block Transformer, FF-dim = 256; 1 block Transformer, FF-dim = 512; 2 block Transformer, FF-dim = 256 và 2 block Transformer, FF-dim = 512. Để đảm bảo tính gọn nhẹ, các biến thể này chỉ giới hạn ở 1 và 2 block.

**Bảng 3**: So sánh độ chính xác trên tập FER2013 của 4 biến thể - SCoTER

| Mô hình | Cấu trúc Transformer | Số tham số | Bộ dữ liệu | Độ chính xác trên tập đánh giá |
|---|---|---|---|---|
| SCoTER Model 1 | 1 block ,FF-dim = 256 | 1,5M | FER2013 | 67% |
| SCoTER Model 2 | 1 block, FF-dim = 512 | 1,7M | FER2013 | 67% |
| SCoTER Model 3 | 2 block FF-dim = 256 | 1,9M | FER2013 | 69% |
| SCoTER Model 4 | 2 block, FF-dim = 512 | 2,2M | FER2013 | 67% |

Dựa vào kết quả Bảng 3 cho thấy hiệu suất của mô hình SCoTER phụ thuộc vào sự điều chỉnh cân bằng giữa chiều sâu kiến trúc (số lượng Block) và năng lực biểu diễn

(kích thước FF-dim). Việc tăng số lượng Transformer Block từ 1 lên 2 khi FF-dim được giữ ở mức 256 đã mang lại sự cải thiện hiệu suất rõ rệt, cụ thể là tăng độ chính xác từ 67% lên 69%. Điều này chứng tỏ việc tích hợp thêm chiều sâu cho phép mô hình học được các phép biểu diễn đa tầng phức tạp hơn, từ đó nâng cao khả năng phân loại cảm xúc. Ngược lại, khi thiết kế mô hình với 2 block với FF-dim = 512 thì độ chính xác trên tập đánh giá không cải thiện (vẫn là 67%), dù số lượng tham số đã mở rộng lên 2.2M. Hiện tượng này cho thấy mô hình trở nên quá phức tạp so với quy mô của tập dữ liệu FER2013, dẫn đến nguy cơ overfitting và làm giảm khả năng khái quát hóa. Do đó, SCoTER Model 3 (2 Block, FF-dim = 256) là mô hình tốt nhất vì đạt được độ chính xác cao nhất (69%) với số tham số được tối ưu (1.9M).

### 4.4 Đánh giá hiệu năng từng khối

Để phân tích đóng góp của từng thành phần trong kiến trúc SCoTER, nhóm nghiên cứu thực hiện thí nghiệm loại trừ bằng cách lần lượt vô hiệu hóa khối Transformer và khối Patch Extraction, kết quả thể hiện ở **Bảng 4**. Cụ thể, khối *Patch Extraction* giúp mô hình nắm bắt đặc trưng cục bộ, trong khi khối *Transformer* mô hình hóa mối quan hệ toàn cục giữa các vùng biểu cảm. Do đó, mô hình đầy đủ được lựa chọn cho các thí nghiệm tiếp theo nhằm đảm bảo sự cân bằng giữa độ chính xác và hiệu quả tính toán.

***Bảng 4.*** Kết quả thí nghiệm loại trừ (Ablation Study) đánh giá hiệu quả của từng khối trong mô hình SCoTER

| Cấu hình mô hình | Patch Extraction | Transformer | Val-Accuracy (FER 2013) |
|---|---|---|---|
| Không Transformer | ✓ | × | 67% |
| Không Patch Extraction | × | ✓ | 67% |
| Mô hình đầy đủ | ✓ | ✓ | 69% |

## 5. KẾT LUẬN

Nghiên cứu đề xuất SCoTER, một mô hình lightweight cải tiến từ PAtt-Lite cho bài toán FER. Các đóng góp chính bao gồm: (1) Tích hợp Multi-Head Attention thay thế Self-Attention đơn, giúp học mối quan hệ toàn cục hiệu quả hơn; (2) Tái thiết kế Patch Extraction với 3 lớp convolution để tăng khả năng biểu diễn cục bộ. Kết quả thực nghiệm cho thấy SCoTER đạt 69% trên FER2013 (tăng 7% so với PAtt-Lite) với chỉ 1.9M tham số, chứng minh hiệu quả trong môi trường thực tế có hạn chế về phần cứng. Thí nghiệm loại trừ khẳng định vai trò bổ trợ của các khối trong việc tăng cường khả năng biểu diễn đặc trưng cảm xúc.Tuy nhiên mô hình vẫn gặp khó khăn với biểu cảm vi tế

và cảm xúc phức hợp. Hiệu năng trên FER2013 (69%) vẫn còn khoảng cách so với state-of-the-art (~75%), do hạn chế về chất lượng dữ liệu và số lượng tham số. Trong tương lai, nhóm nghiên cứu sẽ tiếp tục tối ưu cơ chế attention thích ứng (adaptive attention) nhằm giảm độ phức tạp tính toán, đồng thời mở rộng mô hình cho các bài toán nhận dạng cảm xúc trong môi trường động và đa phương thức.

## TÀI LIỆU THAM KHẢO

[1]. S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Image Process., vol. 29, pp. 5749–5772, 2020.

[2]. W. Wang, H. Zhao, and Y. Wang, "A Survey on Facial Expression Recognition of Static and Dynamic Emotions," arXiv:2408.15777, 2024.

[3]. J. L. Ngwe, K. M. Lim, C. P. Lee, and T. S. Ong, "PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3407108

A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, 2017.

[4]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), 2019.

A. Gursesli, A. L. Ozbay, A. Demir, S. Basaran, B. Gulec, and E. Leloglu, "Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets," in Proc. International Conference on Artificial Intelligence and Data Processing (IDAP), 2023. (fer2013<65)

[5]. G. I. Tutuianu, Y. Liu, A. Alamäki, and J. Kauttonen, "Benchmarking Deep Facial Expression Recognition: An Extensive Protocol with Balanced Dataset in the Wild," *Engineering Applications of Artificial Intelligence*, vol. 136 (108983), pt. B, 2024.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.

[6]. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556, 2014.

[7]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[8]. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar và I. Matthews, "The Extended Cohn-Kanade (CK+) dataset: A complete dataset for action unit and emotion-specified expression," trong *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, San Francisco, CA, USA, 2010, pp. 94-101.

[9]. J. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," *Neural Networks*, vol. 64, pp. 59-63, 2015

# SCoTER: A LIGHTWEIGHT CNN–TRANSFORMER MODEL FOR FACIAL EXPRESSION RECOGNITION

**Tran Thi Kieu[1*], Ho Phuoc Tien[2]**

[1] Faculty of Electronics, Electrical Engineering and Material Technology, University of Sciences, Hue University

[2] Electronics and Telecommunication Engineering, The University of Danang, University of Science and Technology (DUT)

*Email: kieutran@husc.edu.vn

**ABSTRACT**

Facial Expression Recognition (FER) is a fundamental problem in computer vision, playing a vital role in applications such as human–computer interaction, behavior monitoring, and real-time emotion analysis. Although recent deep learning models have made remarkable progress, their high computational complexity remains a major challenge for practical deployment. To address this limitation, this study proposes a lightweight model inspired by PAtt-Lite, aiming to improve recognition accuracy while maintaining computational efficiency. Experimental results demonstrate that the proposed model achieves 100% accuracy on the CK+ dataset and 69% on FER2013 - an improvement of 7% compared to the original model on FER2013 - thus confirming its effectiveness and potential for real-world applications.

**Keywords:** Facial Expression Recognition (FER), MobileNetV1, Transformer, Patch Extraction, Lightweight Model.