

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

CS331.P21 - Thị giác máy tính nâng cao

Trương Huỳnh Thúy An Hoàng Đức Chung Nguyễn Hải Đăng

Khoa Khoa học Máy tính
Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

June 3, 2025

Overview

1. Tổng quan
2. Các nghiên cứu liên quan
3. Phương pháp
4. Thực nghiệm
5. Kết luận

Overview

1. Tổng quan

2. Các nghiên cứu liên quan

3. Phương pháp

4. Thực nghiệm

5. Kết luận

Giới thiệu bài toán

Nhận diện cảm xúc khuôn mặt (Facial Expression Recognition - FER) là một trong những bài toán cốt lõi của thị giác máy tính, với nhiều ứng dụng trong:

- Giao tiếp người – máy (Human-Computer Interaction)
- Phân tích hành vi người dùng (User Behavior Analysis)
- Hệ thống chăm sóc sức khỏe thông minh
- Giám sát và an ninh

Mục tiêu chính của FER là phân loại cảm xúc của con người dựa trên ảnh khuôn mặt, thường bao gồm 7 cảm xúc cơ bản: *Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise*.

Lý do chọn đề tài

Mặc dù các mô hình học sâu hiện đại như CNN, Vision Transformer hay Swin Transformer đã đạt được nhiều thành tựu trong FER, vẫn còn nhiều thách thức tồn tại:

- Thiếu khả năng học đặc trưng cục bộ chi tiết — rất quan trọng với FER.
- Dữ liệu không đồng đều và giới hạn về kích thước, dẫn đến overfitting.
- Thiếu tích hợp hiệu quả giữa các mô hình học toàn cục và cục bộ.

Do đó, nhóm chọn nghiên cứu và đề xuất mô hình lai mới kết hợp Swin Transformer với ResNet và các mô-đun cải tiến như SE để nâng cao hiệu suất nhận diện cảm xúc.

Overview

1. Tổng quan
- 2. Các nghiên cứu liên quan**
3. Phương pháp
4. Thực nghiệm
5. Kết luận

Các nghiên cứu liên quan

TransFER (Xue et al. [1])

- Áp dụng Multi-Attention Dropping (MAD).
- Tập trung vào vùng quan trọng.
- Giảm nhiễu không cần thiết.

VTFF (Ma et al. [5])

- Kết hợp CNN hai nhánh và Transformer.
- Dùng cơ chế ASF (Attentional Selective Fusion).
- Trích xuất thông tin từ đặc trưng “thị giác”.

Swin Transformer [3]

- Self-attention theo cửa sổ trượt.
- Khai thác đặc trưng toàn cục và cục bộ.
- Tăng khả năng biểu diễn hình ảnh.

CT-DBN (Liang et al. [3])

- Kết hợp CNN với Swin Transformer.
- Giải quyết che khuất, thay đổi góc nhìn.

Các nghiên cứu liên quan

SwinT-SE-SAM (Vats & Chadha [2])

Kiến trúc đề xuất

- Swin Transformer + SE block + SAM.
- Tăng hiệu quả nhận diện cảm xúc.

Thực nghiệm

- Dữ liệu: FER2013 (40K), CK+ (1K), AffectNet (60K).
- Tiền xử lý: chuẩn hóa, xoay, thay đổi độ tương phản, augmentation.
- F1-score đạt 0.5420 trên AffectNet (4K ảnh test).

Swin-FER (Bie et al. [4])

Cải tiến chính

- Hợp nhất đặc trưng giữa các tầng.
- Group Convolution.
- Thêm Mean/Split Module.

Kết quả thực nghiệm

- FER2013: 71.11% accuracy.
- CK+: 100% accuracy.

Đề xuất: Swin-SE-ResNet

Kiến trúc

- Swin Transformer + ResNet-50.
- Kết hợp học đặc trưng toàn cục và cục bộ.

Tăng cường:

- Squeeze-and-Excitation (SE block).

Chi tiết kỹ thuật

- Split Convolution và Mean Pooling.
- Giảm nhiễu, giảm tham số dư thừa.
- Huấn luyện trên FER2013 và AffectNet.
- Tăng khả năng khái quát hóa mô hình.

Overview

1. Tổng quan
2. Các nghiên cứu liên quan
- 3. Phương pháp**
4. Thực nghiệm
5. Kết luận

Động lực và lý do lựa chọn

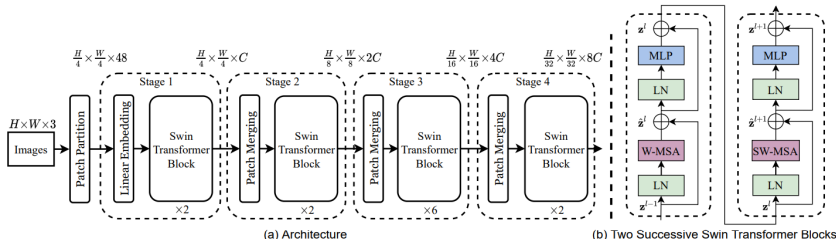
- CNN (VGG, ResNet) hiệu quả trong trích xuất đặc trưng cục bộ nhưng bị giới hạn về receptive field.
- Vision Transformer (ViT) khai thác quan hệ toàn cục qua self-attention, nhưng cần dữ liệu lớn và không tối ưu cho đặc trưng cục bộ.
- **Giải pháp:** Kết hợp Swin Transformer Tiny và ResNet-50 để tận dụng ưu điểm của cả hai.
- **Tăng cường:** Thêm SE block, Split Convolution và Mean Pooling để tăng khả năng biểu diễn và giảm overfitting.

Swin Transformer: Cơ chế Attention hiệu quả

- **Swin Transformer [3]** là kiến trúc Vision Transformer phân cấp, tối ưu cho các tác vụ thị giác.
- Sử dụng **Window-based Multi-head Self-Attention (W-MSA)**:
 - Chia ảnh thành các cửa sổ kích thước cố định ($M \times M$).
 - Attention được tính riêng biệt trong từng cửa sổ \Rightarrow giảm độ phức tạp xuống $\mathcal{O}(M^2 \cdot \frac{N}{M^2})$.
- **Shifted Window Multi-head Self-Attention (SW-MSA)**:
 - Dịch chuyển vị trí các cửa sổ ở tầng kế tiếp.
 - Cho phép mô hình hóa thông tin liên vùng mà vẫn giữ hiệu suất tính toán.

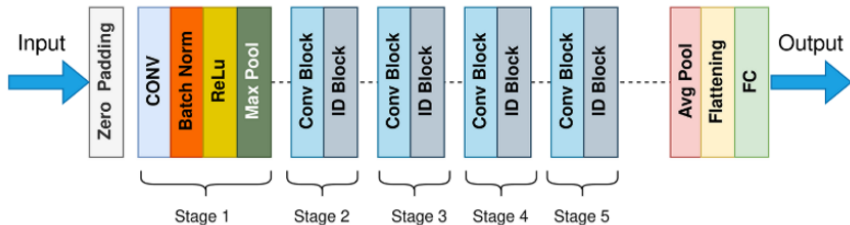
Swin Transformer: Kiến trúc phân cấp

- Swin Transformer gồm **4 giai đoạn chính**:
 1. **Patch Embedding**: Chuyển ảnh đầu vào thành các patch nhỏ (4×4).
 2. **Patch Merging**: Giảm độ phân giải không gian, tăng chiều kênh đặc trưng.
 3. **Swin Transformer Block**: Kết hợp W-MSA và SW-MSA.
 4. **MLP và Residual Connection**: Giúp học biểu diễn mạnh mẽ và ổn định hơn.
- Tạo biểu diễn **phân cấp** giống như CNN, phù hợp với các bài toán phân loại, phát hiện vật thể, phân đoạn,...



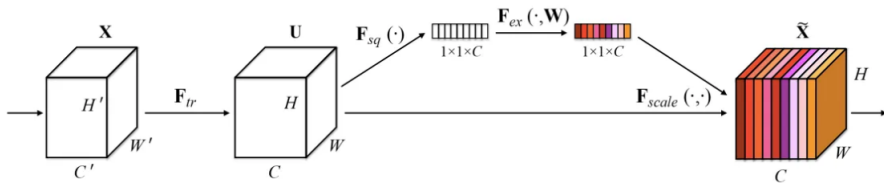
ResNet-50: Backbone học đặc trưng cục bộ

- **ResNet-50 [6]** là mạng CNN gồm 50 tầng, sử dụng **kết nối tắt (Residual Connection)** để tránh mất mát gradient khi huấn luyện sâu.
- Gồm hai loại khối chính:
 - **Identity Block**
 - **Convolutional Block**
- Học tốt đặc trưng cục bộ như mắt, miệng, mũi — những vùng chứa nhiều thông tin biểu cảm.
- Thường được sử dụng như **backbone trích xuất đặc trưng** trong các mô hình nhận diện biểu cảm khuôn mặt.



SE Block: Squeeze-and-Excitation

- Cơ chế attention theo chiều kênh.
- **Squeeze:** Global Average Pooling $\Rightarrow z \in \mathbb{R}^C$.
- **Excitation:** MLP hai lớp \Rightarrow vector trọng số kênh s .
- **Tái hiệu chỉnh:** $\hat{X}_c = s_c \cdot X_c$.



Split Convolution và Mean Pooling

- **Split Convolution:** Chia tensor thành các nhóm kênh, xử lý riêng biệt.
- **Concat:** Ghép lại thành đầu ra tổng hợp.
- **Mean Pooling:** Làm mượt không gian đặc trưng, giảm chiều.

Input

4	7	1	2
6	3	0	8
9	1	6	0
6	4	1	7

Mean
pooling →

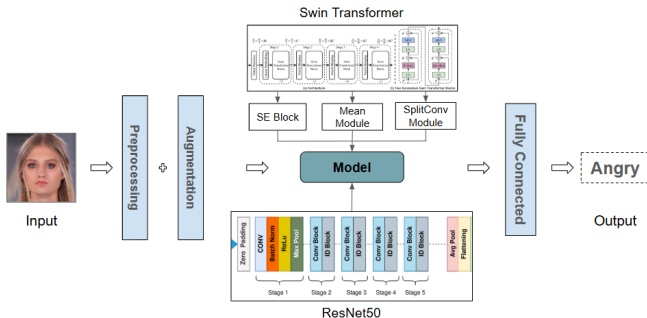
Output

5	2.75
5	3.5

Kiến trúc tổng thể Swin-SE-ResNet

- **Nhánh Swin Transformer:** Trích xuất đặc trưng toàn cục → SE Block → Split Conv → Mean Pooling.
- **Nhánh ResNet-50:** Trích xuất đặc trưng cục bộ → Global Average Pooling.
- **Kết hợp:** Hai vector đầu ra → Fully Connected → Softmax:

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



Overview

1. Tổng quan
2. Các nghiên cứu liên quan
3. Phương pháp
- 4. Thực nghiệm**
5. Kết luận

Bộ dữ liệu sử dụng

Kết hợp hai bộ dữ liệu phổ biến:

- **FER2013**: 35,887 ảnh xám 48×48 , 7 cảm xúc cơ bản.

angry



disgust



fear



happy



neutral



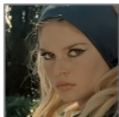
sad



surprise



- **AffectNet**: Lấy mẫu 22,244 ảnh màu 96×96 , có gán nhãn thủ công.



angry



disgust



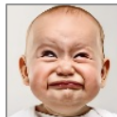
fear



happy



neutral



sad



surprise

Cả hai bộ đều bao gồm 7 nhãn: *Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise*.

Đặc điểm:

- Ảnh xám 48×48 , thu thập từ môi trường thực.
- Không cân bằng lớp cảm xúc.

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

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

Đặc điểm:

- Ảnh màu thu thập từ Internet bằng từ khóa cảm xúc.
- Gán nhãn thủ công cho 7 cảm xúc cơ bản.
- Chọn mẫu gồm 22,244 ảnh kích thước 96×96 .

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

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

Kết hợp FER2013 và AffectNet

Mục tiêu: Tăng đa dạng và cân bằng lớp cảm xúc.

- Gộp ảnh từ 2 bộ theo từng nhãn.
- Lọc ảnh lỗi, đồng bộ kích thước, cân bằng lại số lượng.
- Tổng cộng: **38,550 ảnh**.

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

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

Tiền xử lý và chia dữ liệu

Tiền xử lý:

- Resize về 224×224 , chuyển tensor.
- Chuẩn hóa theo ImageNet mean/std.
- Augment: lật ngang, xoay, tịnh tiến, biến đổi màu.

Chia tập: 8:1:1 cho train, validation và test.

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

Table: Tổng hợp số lượng ảnh theo từng tập giữa các bộ dữ liệu

Huấn luyện trên FER2013

- Early stopping tại epoch 40
- **Train:** Loss = 0.1230, Acc = 95.88%
- **Val:** Loss = 1.3988, Acc = 70.08%, F1 = 70.01%
- **Test:** Acc = 71.64%, F1 = 71.71%

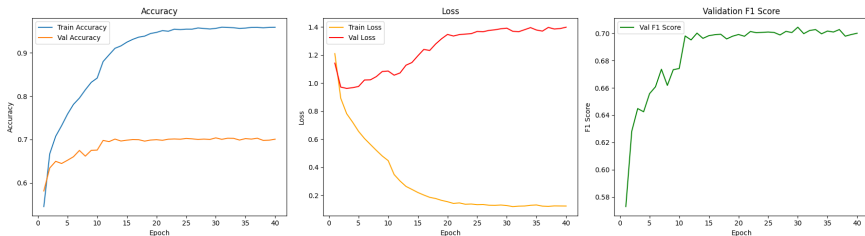


Figure: Quá trình huấn luyện trên FER2013

Confusion Matrix - FER2013

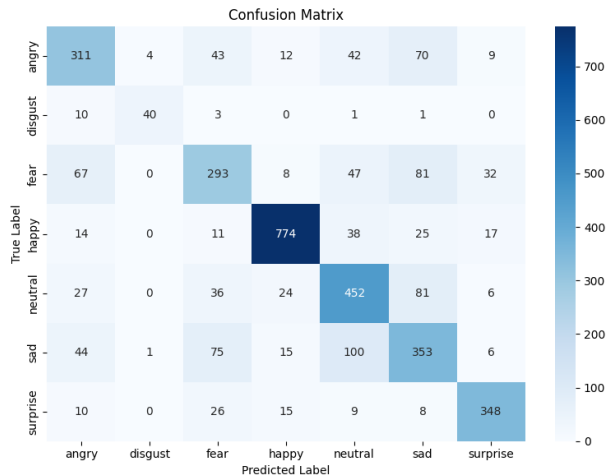


Figure: Confusion Matrix trên tập kiểm tra

Huấn luyện trên AffectNet

- Early stopping tại epoch 22
- **Train:** Loss = 0.0578, Acc = 98.04%
- **Val:** Loss = 0.9684, Acc = 75.79%, F1 = 75.76%
- **Test:** Acc = 76.72%, F1 = 76.81%

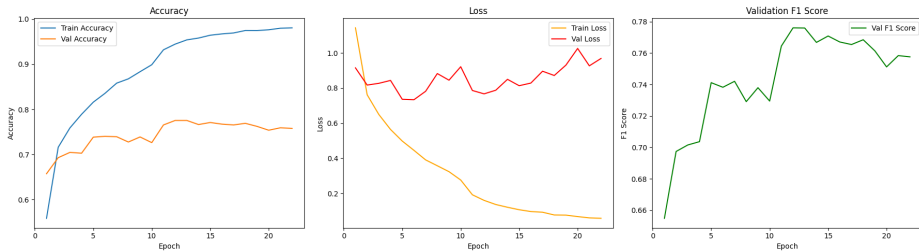


Figure: Quá trình huấn luyện trên AffectNet

Confusion Matrix - AffectNet

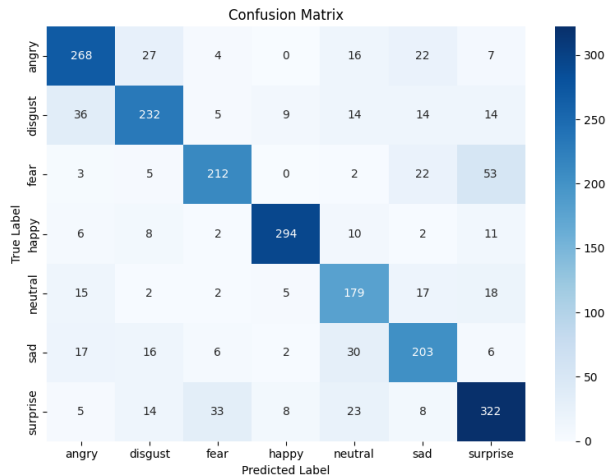


Figure: Confusion Matrix trên tập kiểm tra

Huấn luyện trên FER2013 & AffectNet

- Early stopping tại epoch 38
- **Train:** Loss = 0.0995, Acc = 96.56%
- **Val:** Loss = 1.2269, Acc = 71.86%, F1 = 71.78%
- **Test:** Acc = 73.26%, F1 = 73.24%

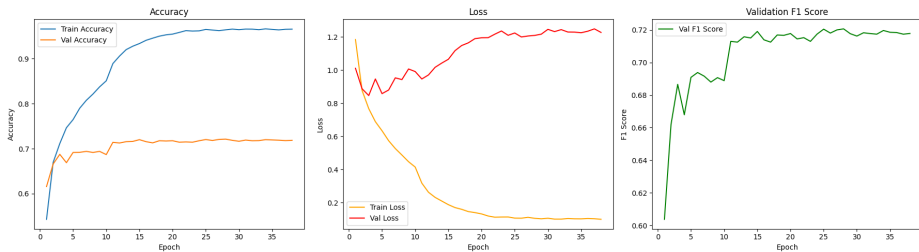


Figure: Quá trình huấn luyện trên FER2013 + AffectNet

Confusion Matrix - FER2013 & AffectNet

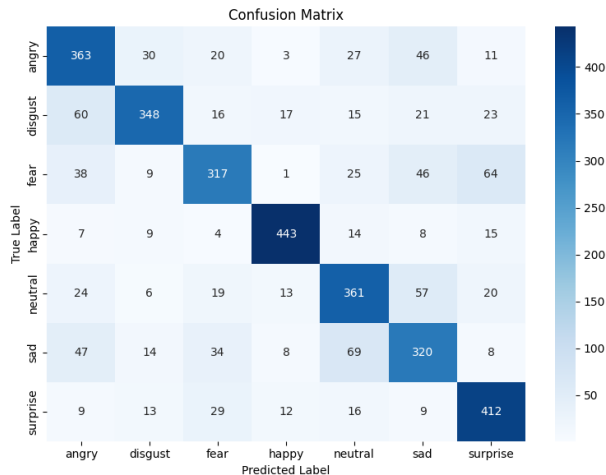


Figure: Confusion Matrix trên tập kiểm tra

So sánh hiệu suất với các mô hình khác trên FER2013

Kết quả trên tập dữ liệu FER2013 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 [6]	66.00%
VGG16+SEBlock [5]	66.80%
Swin-FER [4]	71.11%
Swin-SE-ResNet (Ours)	71.64%

So sánh hiệu suất với các mô hình khác trên AffectNet

Trên tập AffectNet, nhóm chúng tôi 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.

Mô hình	Accuracy	F1-score (Weighted Avg.)
SwinT	76.49%	76.45%
ResNet50	76.63%	76.66%
SwinT-SE-SAM [2]	–	54.20%
TransFER [1]	66.23%	–
Swin-SE-ResNet (Ours)	76.72%	76.81%

- Sử dụng **MTCNN** để phát hiện khuôn mặt trên từng khung hình.
- Khuôn mặt được **resize về 224×224** , chuẩn hóa theo thống kê ImageNet, sau đó đưa vào mô hình.
- Mô hình được **huấn luyện trước trên FER2013** nên có khả năng nhận diện tốt các cảm xúc như *happy*, *sad*, *angry*, *fear*.
- Nhãn cảm xúc được hiển thị trực tiếp trên ảnh cùng với *bounding box*.

Overview

1. Tổng quan
2. Các nghiên cứu liên quan
3. Phương pháp
4. Thực nghiệm
- 5. Kết luận**

Kết luận

- **Swin-SE-ResNet** đạt kết quả tốt trên 3 tập dữ liệu, cao nhất trên **AffectNet**: Accuracy 76.72%.
- Hiệu quả từ việc tích hợp Swin + SE-block + ResNet: trích xuất đặc trưng mạnh mẽ, tập trung vào vùng thông tin quan trọng
- Dữ liệu kết hợp giúp tăng khả năng khái quát hóa: Giảm nhầm lẫn ở các cảm xúc tiêu cực như *fear*, *disgust*, *angry*
- Mô hình hoạt động khá tốt trong môi trường thực tế:
 - Ổn định với dữ liệu liên tục và ngoài phân phối
 - Thích nghi với ánh sáng và đầu vào phức tạp
- **Hạn chế:** Mô hình vẫn có sai sót trong việc phân loại các biểu cảm có biểu hiện gần nhau (như *fear* và *surprise*), đặc biệt trên ảnh nhiễu hoặc biểu cảm mờ.







Hướng phát triển

- Tối ưu mô hình để chạy tốt hơn trên thiết bị di động hoặc hệ thống nhúng.
- Mở rộng dữ liệu huấn luyện đa dạng về **văn hoá, giới tính, độ tuổi**.
- Ứng dụng trong các lĩnh vực:
 - Hỗ trợ tâm lý, chăm sóc sức khoẻ tinh thần.
 - Giao tiếp người–máy (HCI).
- Thử nghiệm các kiến trúc nhẹ: **EfficientFormer, MobileViT, EdgeNeXt**.
- Tích hợp tín hiệu đa modal: **giọng nói, cử chỉ, văn bản**.

Tài liệu tham khảo (1/2)

-  I. Goodfellow et al., *Challenges in Representation Learning: A Report on Three Machine Learning Contests*, ICONIP, 2013.
-  A. Mollahosseini et al., *AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild*, IEEE TAC, 2017.
-  Z. Liu et al., *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*, ICCV, 2021.
-  M. Bie et al., *Swin-FER: Swin Transformer for Facial Expression Recognition*, Applied Sciences, 2024.
-  F. Ma et al., *Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion*, IEEE TAC, 2021.
-  D. Alamsyah and D. Pratama, *Implementasi CNN untuk Klasifikasi Ekspresi Wajah trên FER-2013*, Jurnal Teknologi Informasi, 2020.

Tài liệu tham khảo (2/2)

-  F. Xue et al., *TransFER: Learning Relation-Aware Facial Expression Representations with Transformers*, ICCV, 2021.
-  A. Vats and A. Chadha, *Facial Expression Recognition Using SE-Powered Swin Transformers*, arXiv:2301.10906, 2023.
-  X. Liang et al., *A Convolution–Transformer Dual Branch Network for Facial Expression Recognition*, The Visual Computer, 2023.
-  J. Hu et al., *Squeeze-and-Excitation Networks*, CVPR, 2018.
-  H. Nie, *Face Expression Classification Using SE-based VGG16*, ICCECE, 2022.
-  K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.