UỶ BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

NHẬN DIỆN BIỂU CẢM KHUÔN MẶT TRONG ĐIỀU KIỆN ÁNH SÁNG YẾU SỬ DỤNG CNN NHẸ KẾT HỢP KỸ THUẬT TĂNG CƯỜNG DỮ LIỆU THÍCH ỨNG

LUẬN VĂN MÔN HỌC NCKH TRONG CNTT

NGÀNH: CÔNG NGHỆ THÔNG TIN

Nhóm sinh viên thực hiện:

Họ và tên	MSSV
Văn Tuấn Kiệt	3122410202
Mai Phúc Lâm	3122410207
Nguyễn Đức Duy Lâm	3122410208
Nguyễn Hữu Lộc	3122410213

Giáo viên hướng dẫn: Đỗ Như Tài

TP.HCM, 2025

LỜI CẢM ƠN

Để hoàn thành dự án nghiên cứu khoa học "Nhận diện biểu cảm khuôn mặt trong điều kiện ánh sáng yếu sử dụng CNN nhẹ kết hợp kỹ thuật tăng cường dữ liệu thích ứng" nhóm chúng em đã nhận được rất nhiều sự giúp đỡ và hỗ trợ tận tình từ nhiều phía.

Trước hết, chúng em xin gửi lời cảm ơn chân thành đến Khoa Công Nghệ Thông Tin – Trường Đại Học Sài Gòn đã tạo mọi điều kiện thuận lợi để nhóm có thể thực hiện nghiên cứu này.

Chúng em cũng xin bày tỏ lòng biết ơn sâu sắc đến thầy Nguyễn Quốc Huy và thầy Đỗ Như Tài, những người đã tận tình hướng dẫn, chỉ bảo và hỗ trợ trong suốt quá trình thực hiện đề tài. Sự định hướng và hỗ trợ quý báu của thầy đã giúp nhóm hoàn thành nghiên cứu một cách hiệu quả và thành công.

Ngoài ra, chúng em xin cảm ơn các thành viên trong nhóm đã luôn đoàn kết, hỗ trợ lẫn nhau và nỗ lực hết mình để đưa dự án đến kết quả tốt nhất.

Cuối cùng, chúng em xin kính chúc các thầy cô luôn mạnh khỏe, hạnh phúc và thành công để tiếp tục dìu dắt các thế hệ học sinh, sinh viên trên con đường học tập và nghiên cứu.

TÓM TẮT

Nhận diện biểu cảm khuôn mặt (Facial Expression Recognition - FER) trong điều kiện ánh sáng yếu là một thách thức lớn của thị giác máy tính, đặc biệt trên các thiết bị nhúng với tài nguyên hạn chế.

Nghiên cứu này đề xuất một phương pháp hiệu quả kết hợp mạng nơ-ron tích chập nhẹ MobileNetV3 với kỹ thuật tăng cường dữ liệu thích ứng, nhằm nâng cao hiệu suất FER trong điều kiện ánh sáng yếu. Hệ thống xử lý ảnh được thiết kế để tự động lựa chọn các kỹ thuật tiền xử lý (gamma correction, CLAHE, contrast stretching) dựa trên đặc trưng ánh sáng đầu vào.

Thí nghiệm trên tập dữ liệu FER-2013 với ảnh ánh sáng yếu mô phỏng cho thấy mô hình đạt độ chính xác 61.55%, gần tương đương với hiệu suất trên tập gốc (61.63%). Trong khi đó, khi chỉ áp dụng phép giảm độ sáng, độ chính xác giảm còn 58.86%. Như vậy, thuật toán đề xuất giúp cải thiện gần 2.7%, chứng tỏ tính hiệu quả trong việc tăng độ bền mô hình trước biến thiên ánh sáng. Đối chiếu với ResNet18 (đạt 67.48%), MobileNetV3 vẫn cho thấy ưu thế về tốc độ và kích thước mô hình.

Phương pháp này không chỉ cải thiện độ chính xác trong môi trường ánh sáng yếu mà còn phù hợp cho các ứng dụng thực tế như camera giám sát hoặc thiết bị IoT. Nghiên cứu cũng mở ra hướng phát triển cho các giải pháp FER nhẹ, thích ứng và hiệu quả hơn trong tương lai.

1 Tổng quan vấn đề

1.1 Lý do chọn đề tài

Nhận diện biểu cảm khuôn mặt (Facial Expression Recognition - FER) đóng vai trò quan trọng trong các ứng dụng thực tiễn như giao tiếp người-máy, giám sát an ninh, và phân tích hành vi. Tuy nhiên, trong các điều kiện ánh sáng yếu, chẳng hạn như môi trường ban đêm hoặc khu vực thiếu sáng, hiệu quả của các hệ thống FER giảm đáng kể do chất lượng hình ảnh thấp. Các nghiên cứu gần đây (2020–2025) chủ yếu tập trung vào điều kiện ánh sáng lý tưởng, trong khi các giải pháp cho ánh sáng yếu thường phức tạp, đòi hỏi tài nguyên tính toán lớn hoặc không tối ưu cho các thiết bị nhúng.

Việc phát triển một phương pháp nhận diện biểu cảm hiệu quả trong điều kiện ánh sáng yếu, sử dụng mô hình CNN nhẹ (như MobileNetV3) và kỹ thuật tăng cường dữ liệu thích ứng, không chỉ đáp ứng nhu cầu thực tiễn mà còn mang lại giá trị khoa học thông qua việc cải tiến các kỹ thuật hiện có. Đề tài này được chọn vì tính khả thi trong thời gian nghiên cứu (6 tuần), tính mới trong việc kết hợp các phương pháp đơn giản nhưng hiệu quả, và tiềm năng ứng dụng trong các hệ thống thực tế như camera giám sát hoặc thiết bị IoT.

1.2 Vấn đề nghiên cứu

Trong điều kiện ánh sáng yếu, các mô hình nhận diện biểu cảm khuôn mặt truyền thống thường gặp khó khăn do độ tương phản thấp, nhiễu ảnh, và mất chi tiết khuôn mặt. Các phương pháp hiện tại như sử dụng GAN (Generative Adversarial Networks) hoặc Retinex-based preprocessing tuy hiệu quả nhưng phức tạp, yêu cầu thời gian huấn luyện lâu và tài nguyên tính toán lớn, không phù hợp với các ứng dụng thời gian thực hoặc thiết bị có tài nguyên hạn chế. Ngoài ra, các kỹ thuật tăng cường dữ liệu cố định (fixed augmentation) không tối ưu vì không thích nghi với mức độ ánh sáng yếu khác nhau của từng ảnh.

Vấn đề nghiên cứu được đặt ra là: Làm thế nào để phát triển một hệ thống nhận diện biểu cảm khuôn mặt trong điều kiện ánh sáng yếu, sử dụng mô hình CNN nhẹ và kỹ thuật tăng cường dữ liệu thích ứng, nhằm đạt được độ chính xác

cao, tốc độ xử lý nhanh, và khả năng triển khai trên các thiết bị nhúng?

1.3 Mục tiêu nghiên cứu

Mục tiêu tổng quát của nghiên cứu là xây dựng một hệ thống nhận diện biểu cảm khuôn mặt hiệu quả trong điều kiện ánh sáng yếu, sử dụng mạng nơ-ron tích chập nhẹ (MobileNetV3) kết hợp với kỹ thuật tăng cường dữ liệu thích ứng. Các mục tiêu cụ thể bao gồm:

1. Phát triển một pipeline tăng cường dữ liệu thích ứng, tự động điều chỉnh các kỹ thuật tăng cường dựa trên mức độ ánh sáng yếu của từng ảnh.
2. Huấn luyện và tinh chỉnh mô hình MobileNetV3 để nhận diện biểu cảm khuôn mặt trong điều kiện ánh sáng yếu với độ chính xác cao.
3. Đánh giá và so sánh hiệu quả của phương pháp đề xuất với các kỹ thuật tăng cường dữ liệu cố định và các mô hình CNN khác (nếu khả thi).

1.4 Câu hỏi nghiên cứu

Nghiên cứu tập trung trả lời các câu hỏi sau:

1. Làm thế nào để thiết kế một pipeline tăng cường dữ liệu thích ứng, hiệu quả trong việc cải thiện chất lượng ảnh ánh sáng yếu cho nhận diện biểu cảm khuôn mặt?
2. Mô hình MobileNetV3 có thể đạt được độ chính xác tương đương hoặc vượt trội so với các kỹ thuật tăng cường dữ liệu cố định trong điều kiện ánh sáng yếu không?
3. Các kỹ thuật tăng cường dữ liệu thích ứng ảnh hưởng như thế nào đến hiệu suất của mô hình CNN nhẹ trong nhận diện biểu cảm khuôn mặt?

1.5 Phạm vi nghiên cứu

- Đối tượng nghiên cứu: Các biểu cảm khuôn mặt (ví dụ: vui, buồn, tức giận, ngạc nhiên) trong điều kiện ánh sáng yếu, được mô phỏng hoặc thu thập từ bộ dữ liệu công khai FER-2013.
- Phạm vi không gian: Nghiên cứu tập trung vào xử lý hình ảnh tĩnh (static

images), không bao gồm dữ liệu video hoặc dữ liệu đa phổ.

- Phạm vi thời gian: Nghiên cứu được thực hiện trong 8 tuần, từ tháng 4 đến tháng 5 năm 2025, với các thí nghiệm dựa trên dữ liệu công khai và mô hình pre-trained.
- Phạm vi kỹ thuật: Sử dụng mô hình CNN nhẹ (MobileNetV3) và các kỹ thuật tăng cường dữ liệu như gamma correction, histogram equalization, được triển khai bằng Python với các thư viện TensorFlow/Keras và OpenCV.

2 Lược khảo tài liệu

2.1 Tổng hợp các tài liệu, nghiên cứu trước liên quan

2.1.1 Nghiên cứu về nhận diện biểu cảm khuôn mặt (FER)

Nhận diện biểu cảm khuôn mặt (Facial Expression Recognition - FER) là một lĩnh vực trọng điểm trong thị giác máy tính. Từ đầu những năm 2000, các phương pháp truyền thống như LBP, HOG, hoặc SIFT kết hợp với SVM từng chiếm ưu thế. Tuy nhiên, chúng không hiệu quả trong điều kiện ánh sáng thay đổi hoặc góc nhìn khác nhau. Từ năm 2014, học sâu - đặc biệt là CNN - đã nâng cao độ chính xác mô hình FER. Các kiến trúc như VGGNet, ResNet, InceptionNet đạt độ chính xác 70–75% trên FER-2013 nhưng yêu cầu tài nguyên tính toán lớn. [3]

2.1.2 Ảnh hưởng của điều kiện ánh sáng yếu

Zhang et al. (2019), Wang et al. (2022) đã chứng minh rằng ảnh thiếu sáng làm giảm hiệu quả của mô hình FER. GAN-based như EnlightenGAN hoặc RetinexNet giúp cải thiện nhưng đòi hỏi GPU mạnh, không phù hợp với thiết bị thực tế như điện thoại hoặc camera nhúng. [[2], [8]]

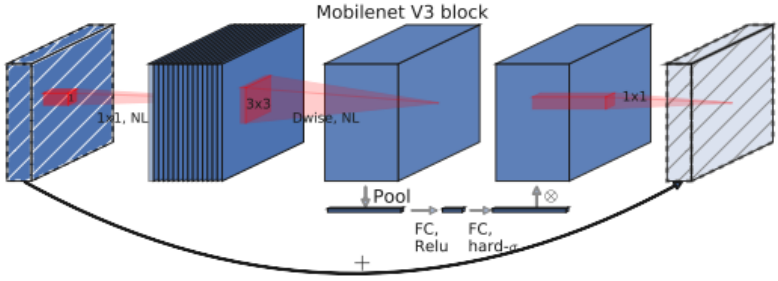
2.1.3 Các kỹ thuật tiền xử lý ảnh tăng cường sáng

Gamma Correction: điều chỉnh độ sáng theo hàm $I_{out} = I_{in}^{\gamma}$. Với $\gamma < 1$, ảnh được làm sáng. [9]

CLAHE: nâng cao độ tương phản cục bộ, phù hợp ảnh có vùng sáng tối không đều. Wang et al. (2021) dùng CLAHE trước FER đạt cải thiện đáng kể. [4]

2.1.4 Mô hình học sâu nhẹ: MobileNetV3

MobileNetV3 (Howard et al., 2019) là CNN nhẹ, tối ưu cho thiết bị di động, gồm kỹ thuật như depthwise separable convolution, squeeze-and-excitation và NAS. MobileNetV3-Small có 2.5M tham số, cân bằng tốt giữa độ chính xác và hiệu suất, chưa được nghiên cứu sâu trong FER ánh sáng yếu. [[5], [6], [7]]



Hình 1: Mô hình MobileNetV3.

2.2 Cơ sở lý thuyết của nghiên cứu

Nghiên cứu kết hợp tiền xử lý ảnh thích ứng theo điều kiện ánh sáng và mô hình CNN nhẹ - MobileNetV3 để tăng hiệu suất FER trong điều kiện ánh sáng yếu.

2.2.1 Tiền xử lý ảnh trong điều kiện ánh sáng yếu

Gamma Correction [9]: hàm phi tuyến giúp làm sáng ảnh thiếu sáng. Ying et al. (2017) chỉ ra rằng γ phù hợp có thể nâng cao chất lượng ảnh mà không gây nhiễu.

CLAHE [4]: phân tích cục bộ từng vùng ảnh, cải thiện chi tiết biểu cảm ở vùng mắt, miệng (Zhu et al., 2018).

Tính thích ứng [10]: thuật toán tự động phân tích histogram và độ sáng trung bình để chọn phương pháp phù hợp (Chen et al., 2021).

2.2.2 Nhận diện biểu cảm bằng mô hình CNN nhẹ – MobileNetV3

MobileNetV3-Small [5] (Howard et al., 2019): 2.5 triệu tham số, thích hợp cho thiết bị nhúng. Nghiên cứu dùng mô hình này để fine-tune phân loại 7 biểu cảm.

Kỹ thuật chính: Depthwise Separable Convolution (Howard et al., 2017) [5], SE Module (Hu et al., 2018) [11], Hard-Swish Activation.

2.2.3 Pipeline đề xuất trong nghiên cứu

Dựa trên hai thành phần lý thuyết đã trình bày, nghiên cứu đề xuất pipeline xử lý gồm 3 giai đoạn chính như trong Bảng 1:

Bảng 1: Pipeline đề xuất trong nghiên cứu

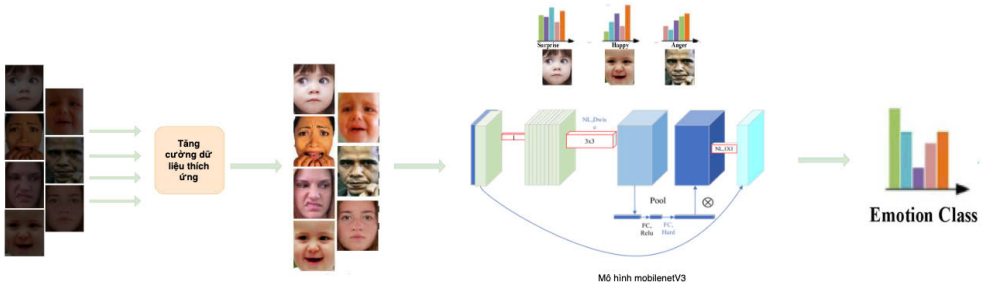
Giai đoạn	Nội dung
Tiền xử lý ảnh	• Chuyển ảnh sang ảnh grayscale. • Tính độ sáng trung bình μ. • Nếu $\mu < T_1$: áp dụng gamma correction với $\gamma \in [0.4, 0.5]$. • Nếu $T_1 < \mu < T_2$: áp dụng CLAHE. • Nếu $\mu > T_2$: giữ nguyên hoặc áp dụng contrast stretching nhẹ.
Học biểu cảm	• Ảnh sau tiền xử lý được đưa vào mô hình MobileNetV3-Small. • Mô hình được fine-tune để phân loại 7 biểu cảm: vui, buồn, giận, sợ, bất ngờ, ghê tởm, trung tính.
Đánh giá mô hình	• Thực hiện trên tập test có và không áp dụng tăng cường ảnh. • Sử dụng các chỉ số đánh giá: – Accuracy – Precision / Recall – F1-score – Confusion Matrix • So sánh với baseline không áp dụng tăng cường để đánh giá hiệu quả thực sự.

Bảng 1 mô tả chi tiết pipeline xử lý được đề xuất trong nghiên cứu, bao gồm ba giai đoạn chính: tiền xử lý ảnh, học biểu cảm, và đánh giá mô hình. Trong giai đoạn tiền xử lý, ảnh đầu vào được chuyển sang ảnh xám và điều chỉnh độ sáng hoặc tương phản dựa trên giá trị trung bình μ của ảnh. Tùy theo mức độ sáng, các kỹ thuật như gamma correction, CLAHE hoặc contrast stretching nhẹ sẽ được áp dụng nhằm cải thiện chất lượng ảnh đầu vào.

Tiếp theo, giai đoạn học biểu cảm sử dụng mô hình MobileNetV3-Small đã được tinh chỉnh (fine-tune) để phân loại bảy loại biểu cảm khuôn mặt phổ biến: vui, buồn, giận, sợ, bất ngờ, ghê tởm và trung tính.

Cuối cùng, mô hình được đánh giá trên tập kiểm thử với và không có áp

dụng kỹ thuật tăng cường ảnh. Các chỉ số đánh giá bao gồm Accuracy, Precision, Recall, F1-score và ma trận nhầm lẫn (Confusion Matrix). Kết quả mô hình sẽ được so sánh với một mô hình baseline không áp dụng tăng cường ảnh nhằm đánh giá hiệu quả thực sự của pipeline đề xuất.



Hình 2: Pine line đề xuất

Hình 2 cho thấy pipeline đề xuất cho hệ thống nhận diện cảm xúc khuôn mặt. Dữ liệu ảnh đầu vào được tăng cường bằng cách điều chỉnh độ sáng nhằm cải thiện khả năng nhận diện trong các điều kiện ánh sáng khác nhau. Sau đó, ảnh được đưa vào mô hình MobileNetV3 để trích xuất đặc trưng và phân loại cảm xúc đầu ra.

2.3 Phân tích điểm mạnh, điểm yếu của các nghiên cứu trước và hướng kế thừa

2.3.1 Điểm mạnh

- Mô hình học sâu giúp tăng độ chính xác FER (Mollahosseini et al., 2016) [3].
- MobileNetV3 hiệu quả, tiết kiệm tài nguyên (Howard et al., 2019) [[5], [6]].
- CLAHE giúp tăng sáng hiệu quả, đơn giản (Wang et al., 2020) [4].

2.3.2 Hạn chế

- Chưa chú trọng ánh sáng yếu trong FER (Barsoum et al., 2016) [[3], [8]].
- Pipeline thiếu bước tăng cường ảnh (Zhou et al., 2021) [12].
- Dùng GAN tăng sáng gây tốn tài nguyên (Chen et al., 2020) [2].

2.3.3 Hướng kế thừa và phát triển

- Chọn MobileNetV3-Small làm backbone (Howard et al., 2019) [5].
- Thiết kế pipeline có bước xử lý ảnh thích ứng đầu vào. [10]

- Mô phỏng tập FER-2013 thiếu sáng để kiểm thử.
- Ưu tiên tăng sáng đơn giản thay vì GAN.

2.4 Cơ sở lý thuyết của thuật toán tăng cường dữ liệu thích ứng

2.4.1 Lý do phát triển thuật toán

Trong bài toán nhận diện biểu cảm khuôn mặt (FER – Facial Expression Recognition) dưới điều kiện ánh sáng yếu, hình ảnh khuôn mặt thường bị suy giảm chất lượng nghiêm trọng do hiện tượng thiếu sáng toàn cục hoặc cục bộ. Điều này dẫn đến hiện tượng mất chi tiết, đặc biệt ở các vùng chứa đặc trưng biểu cảm quan trọng như mắt, miệng, nếp nhăn. Kết quả là mô hình học sâu, vốn phụ thuộc vào độ tương phản và cấu trúc cục bộ, sẽ khó khăn trong việc nhận dạng chính xác. [[10], [12]]

Các kỹ thuật tăng cường dữ liệu truyền thống như histogram equalization hoặc gamma correction thường được áp dụng đồng loạt cho toàn bộ dữ liệu huấn luyện. Tuy nhiên, cách tiếp cận này bỏ qua tính biến thiên về mức sáng của từng ảnh đầu vào. Cụ thể:

- Với ảnh quá tối, tăng sáng quá mức dễ làm mất chi tiết do bão hòa điểm ảnh.
- Với ảnh sáng vừa đủ, tăng cường không cần thiết có thể làm biến dạng đặc trưng tự nhiên, dẫn đến suy giảm hiệu quả học.

Do đó, nghiên cứu này đề xuất một thuật toán tăng cường dữ liệu thích ứng, có khả năng phân tích đặc trưng ánh sáng riêng của từng ảnh, từ đó lựa chọn kỹ thuật xử lý phù hợp, đơn giản nhưng hiệu quả và phù hợp để huấn luyện với mô hình nhẹ như MobileNetV3-Small.

2.4.2 Các thành phần lý thuyết chính

(a) Phân tích độ sáng của ảnh [9]

Để xác định ảnh đầu vào có cần tăng cường hay không, và nếu cần thì sử dụng phương pháp nào, cần phân tích một số đặc trưng cơ bản về độ sáng:

- Độ sáng trung bình (mean intensity): Được tính trên ảnh chuyển sang thang xám (grayscale) hoặc kênh Y (luminance) trong không gian YUV.

$$\mu = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} I(i,j)$$

- Độ lệch chuẩn (standard deviation): Đánh giá mức độ phân tán sáng tối, cho biết ảnh có sáng đồng đều hay có vùng sáng – vùng tối xen kẽ.
- Histogram phân bố pixel: Dùng để xác định ảnh có độ tương phản thấp (hẹp histogram) hoặc bị lệch về vùng tối.

(b) Các kỹ thuật tăng cường ánh sáng được sử dụng [[9], [12]]

- Gamma Correction:

$$I_{\text{out}} = I_{\text{in}}^{\gamma}$$

  - $\gamma < 1$: ảnh được làm sáng lên.
  - $\gamma > 1$: ảnh bị làm tối hơn.

  Việc chọn giá trị γ được tính toán dựa trên giá trị độ sáng trung bình μ của ảnh.

- Histogram Equalization (HE): Phân bố lại giá trị pixel để làm tăng độ tương phản tổng thể. Phù hợp khi histogram bị tập trung ở vùng tối (low dynamic range). Tuy nhiên, dễ gây nhiễu ở ảnh có noise.
- Contrast Stretching: Kéo dãn mức độ sáng từ dải cường độ cũ về dải chuẩn 0–255:

$$I_{\text{out}} = \frac{I_{\text{in}} - I_{\text{min}}}{I_{\text{max}} - I_{\text{min}}} \times 255$$

```python
import cv2
import numpy as np
import matplotlib.pyplot as plt

def adaptive_augmentation(image, T1=50, T2=100, gamma_low=0.5, gamma_mid=0.8):
    gray = cv2.cvtColor(image, cv2.COLOR_RGB2GRAY)
    mean_intensity = np.mean(gray)
    print("T =", mean_intensity)

    if mean_intensity < T1:
        gamma = gamma_low
        image = gamma_correction(image, gamma)
    elif T1 <= mean_intensity < T2:
        gamma = gamma_mid
        image = gamma_correction(image, gamma)
    else:
        image = contrast_stretching(image)

    return image

def gamma_correction(image, gamma):
    table = np.array([((i / 255.0) ** gamma) * 255 for i in np.arange(256)]).astype("uint8")
    return cv2.LUT(image, table)

def histogram_equalization(image):
    img_yuv = cv2.cvtColor(image, cv2.COLOR_RGB2YUV)
    img_yuv[:,:,0] = cv2.equalizeHist(img_yuv[:,:,0])
    return cv2.cvtColor(img_yuv, cv2.COLOR_YUV2RGB)

def contrast_stretching(image):
    min_val = np.min(image)
    max_val = np.max(image)
    stretched = (image - min_val) * 255.0 / (max_val - min_val + 1e-6)
    return stretched.astype(np.uint8)



if __name__ == "__main__":
    # Đọc ảnh đầu vào
    img = cv2.imread("dark_image.png")
    img_rgb = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)

    # Áp dụng tăng cường
    enhanced_img = adaptive_augmentation(img_rgb)

    # Hiển thị ảnh gốc và ảnh tăng cường nằm ngang
    plt.figure(figsize=(10, 5))

    # Ảnh gốc
    plt.subplot(1, 2, 1)
    plt.imshow(img_rgb)
    plt.title("Ảnh gốc (Tối)")
    plt.axis('off')

    # Ảnh sau tăng cường
    plt.subplot(1, 2, 2)
    plt.imshow(enhanced_img)
    plt.title("Ảnh sau tăng cường")
    plt.axis('off')

    plt.tight_layout()
    plt.show()

    cv2.imwrite("enhanced_image.jpg", cv2.cvtColor(enhanced_img, cv2.COLOR_RGB2BGR))
```

Hình 3: Code minh họa

Hình 4: Ảnh trước và sau khi tăng cường

(c) Tính thích ứng của thuật toán [10]

Thuật toán sẽ:

- Tính toán độ sáng trung bình (μ) và độ lệch chuẩn (σ) của từng ảnh đầu vào.
- Dựa vào hai ngưỡng xác định trước T_1 và T_2, phân loại mức độ ánh sáng:
  - $\mu < T_1$ (ảnh rất tối): áp dụng gamma nhỏ (0.3–0.5).
  - $T_1 \leq \mu < T_2$ (tối vừa): áp dụng gamma nhẹ (0.7–0.9) hoặc HE.
  - $\mu \geq T_2$ (sáng đủ): không tăng cường hoặc chỉ contrast stretching nhẹ.

Cách tiếp cận này giúp mỗi ảnh được tăng cường đúng mức, tránh làm hỏng đặc trưng gốc hoặc gây dư sáng.

2.4.3 Nguồn cảm hứng và các nghiên cứu liên quan

Retinex-based methods (Fu et al., 2020) đề xuất kỹ thuật phân tách ảnh thành hai thành phần: phản xạ và ánh sáng chiếu vào, sau đó tái cấu trúc lại ảnh với độ sáng cải thiện. Phương pháp này cho kết quả nâng cao rõ rệt nhưng đòi hỏi thuật toán phức tạp và tài nguyên tính toán lớn, do đó khó triển khai trên các thiết bị nhúng. [2]

GAN-based methods như EnlightenGAN (Jiang et al., 2019) sử dụng mạng sinh ảnh để tạo lại phiên bản ảnh có ánh sáng tốt hơn từ ảnh thiếu sáng ban đầu. Mặc dù đem lại chất lượng thị giác cao, nhưng các mô hình GAN thường yêu cầu

GPU mạnh và thời gian xử lý lâu, khiến chúng không phù hợp với các ứng dụng thời gian thực trên thiết bị di động. [2]

Adaptive Augmentation trong học sâu (Zhang et al., 2021) nhấn mạnh tầm quan trọng của việc sử dụng đặc trưng đầu vào để quyết định chiến lược tăng cường dữ liệu phù hợp, thay vì áp dụng cố định một kỹ thuật như truyền thống. Điều này giúp mô hình học sâu đạt hiệu quả tốt hơn trong môi trường đầu vào đa dạng.

Từ các nghiên cứu trên, thuật toán của nhóm đề xuất kế thừa ý tưởng adaptive preprocessing, nhưng được đơn giản hóa để giảm chi phí tính toán và đảm bảo tính linh hoạt, phù hợp với các mô hình nhẹ như MobileNetV3. [12]

3 Phương pháp nghiên cứu

3.1 Thiết kế nghiên cứu

Nghiên cứu được thiết kế theo phương pháp định lượng, tập trung vào việc xây dựng và đánh giá hiệu suất của các mô hình học sâu trong bài toán nhận diện biểu cảm khuôn mặt (Facial Expression Recognition - FER) trong điều kiện ánh sáng yếu. Phương pháp định lượng được chọn vì mục tiêu nghiên cứu là đo lường các chỉ số hiệu suất cụ thể (Accuracy, Precision, Recall, F1-score và thời gian suy luận) của hai mô hình CNN: MobileNetV3 (mô hình nhẹ) và ResNet18 (mô hình sâu hơn), khi áp dụng kỹ thuật tăng cường dữ liệu thích ứng.

Quá trình nghiên cứu bao gồm ba giai đoạn chính:

- Tiền xử lý dữ liệu: Sử dụng tập dữ liệu FER-2013, áp dụng các kỹ thuật tăng cường dữ liệu thích ứng để mô phỏng điều kiện ánh sáng yếu.
- Huấn luyện và tối ưu mô hình: Triển khai MobileNetV3 và ResNet18, tinh chỉnh các tham số để phù hợp với bài toán FER.
- Đánh giá và so sánh: So sánh hiệu suất và thời gian suy luận của các mô hình khi có và không áp dụng kỹ thuật tăng cường dữ liệu thích ứng.

3.2 Đối tượng và mẫu nghiên cứu

3.2.1 Đối tượng nghiên cứu

Đối tượng nghiên cứu là các kỹ thuật nhận diện biểu cảm khuôn mặt trong điều kiện ánh sáng yếu, với trọng tâm vào:

- Mô hình học sâu: MobileNetV3 và ResNet18 dùng để phân loại 7 biểu cảm khuôn mặt (vui, buồn, tức giận, sợ hãi, ngạc nhiên, ghê tởm, trung lập).
- Kỹ thuật tăng cường dữ liệu thích ứng: Các phương pháp như gamma correction, contrast stretching và histogram equalization, được điều chỉnh dựa trên đặc trưng ánh sáng của hình ảnh.

3.2.2 Mẫu nghiên cứu

Mẫu nghiên cứu là tập dữ liệu FER-2013, chứa 35.887 hình ảnh khuôn mặt (48x48 pixel, ảnh xám) được phân loại thành 7 biểu cảm. Tập dữ liệu được chia như sau:

- Tập huấn luyện: 28.709 hình ảnh (80%).
- Tập xác thực (validation): 3.589 hình ảnh (10.00%).
- Tập kiểm tra: 3.589 hình ảnh (10.00%).

Nhằm mô phỏng điều kiện ánh sáng yếu, một tập dữ liệu phụ được tạo ra bằng cách giảm độ sáng của ảnh gốc. Quá trình này thực hiện bằng cách chuyển ảnh sang không gian màu HSV, giảm kênh độ sáng (V) theo một hệ số cố định, sau đó chuyển lại về không gian RGB. Cụ thể, độ sáng được giảm xuống 10% so với ảnh ban đầu.

3.3 Cách thu thập dữ liệu

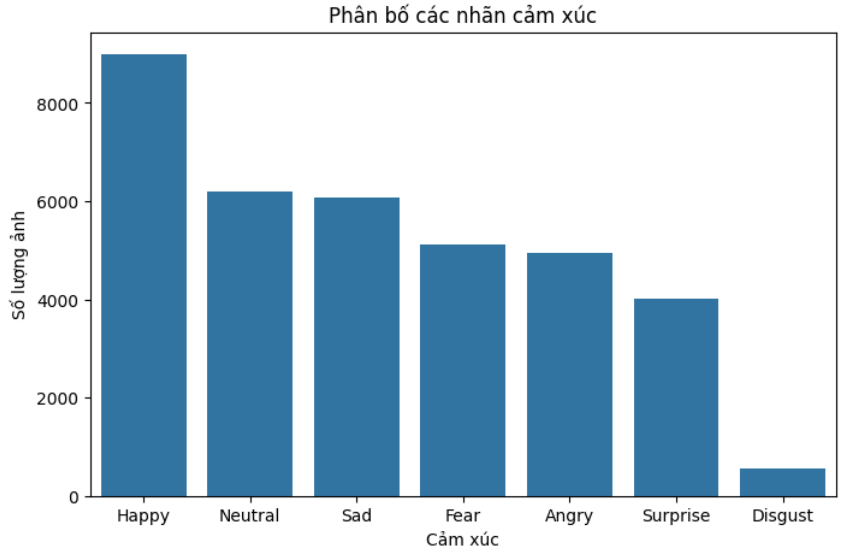
Dữ liệu được thu thập từ tập dữ liệu FER-2013 công khai trên nền tảng Kaggle. Các bước gồm:

Thu thập dữ liệu

- Tải tập dữ liệu FER-2013 từ Kaggle.
- Kiểm tra tính toàn vẹn (số lượng ảnh, định dạng, chất lượng).

3.3.1 EDA dữ liệu

Phân tích dữ liệu khám phá (EDA) được thực hiện trên tập dữ liệu FER-2013 nhằm hiểu rõ cấu trúc, phân phối và đặc trưng của dữ liệu trước khi áp dụng các mô hình học sâu. Dữ liệu được lưu trữ dưới dạng tệp CSV với ba cột chính: cột emotion (nhãn cảm xúc, giá trị từ 0 đến 6), cột pixels (tập hợp các giá trị pixel của ảnh dưới dạng chuỗi số), và cột Usage (chỉ định tập huấn luyện, xác thực hoặc kiểm tra). Kích thước tổng cộng của tập dữ liệu là 35.887 mẫu, trong đó mỗi hình ảnh có độ phân giải 48x48 pixel.

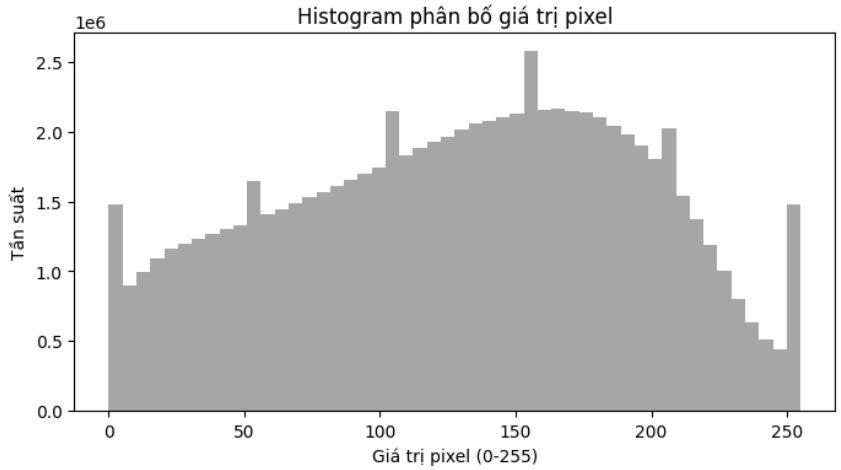


Hình 5: Phân bố nhãn cảm xúc trong tập dữ liệu FER-2013.

Biểu đồ hình 5 thể hiện phân bố các nhãn cảm xúc trong tập dữ liệu FER-2013. Có thể thấy, nhãn "Happy" chiếm tỷ lệ lớn nhất với số lượng ảnh vượt trội so với các nhãn khác, cho thấy sự mất cân bằng dữ liệu. Các cảm xúc như "Neutral", "Sad", "Fear" và "Angry" có số lượng tương đối đồng đều, trong khi "Surprise" ít hơn và "Disgust" là cảm xúc có số lượng ảnh ít nhất. Điều này cho thấy việc huấn luyện mô hình nhận diện cảm xúc dựa trên tập dữ liệu này có thể gặp khó khăn với các nhãn có số lượng ít, đặc biệt là "Disgust", do thiếu dữ liệu đại diện.

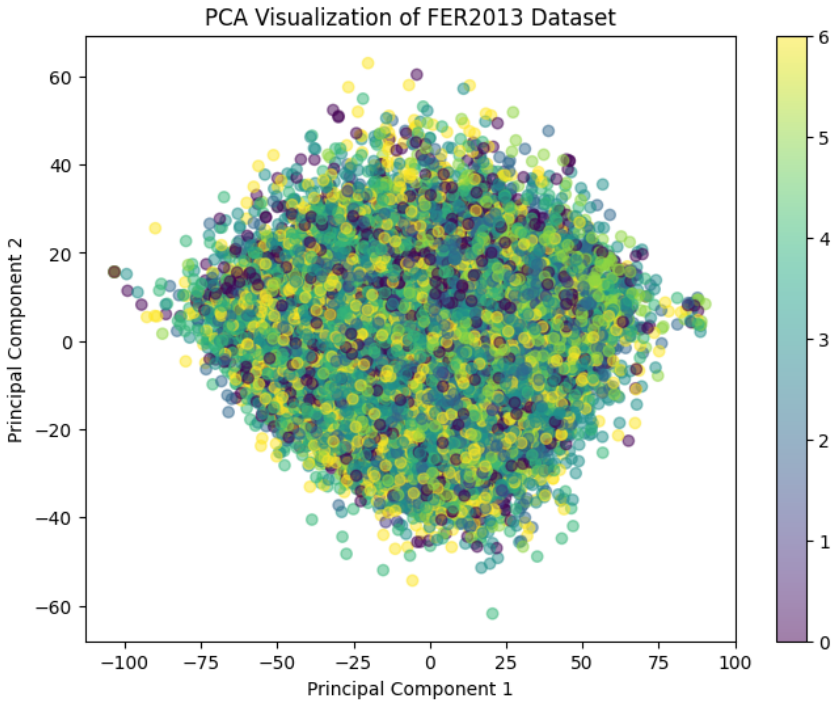
Hình 6: Một số hình ảnh mẫu từ tập dữ liệu FER-2013.

Hình 6 cho thấy tập dữ liệu có rất nhiều biến thể về tuổi tác, góc độ, cường độ biểu cảm, và sự xuất hiện trong môi trường thực tế



Hình 7: Phân bố pixel trong tập dữ liệu FER-2013.

Biểu đồ hình 7 mô tả phân bố giá trị pixel trong tập dữ liệu FER-2013. Trục hoành thể hiện các giá trị pixel từ 0 đến 255, trong khi trục tung thể hiện tần suất xuất hiện tương ứng. Từ biểu đồ có thể thấy rằng các giá trị pixel được phân bố tương đối rộng và không đồng đều, với một số đỉnh rõ rệt tại các giá trị như 0, 100, 150 và 255. Điều này cho thấy dữ liệu ảnh trong FER-2013 có sự đa dạng về độ sáng và độ tương phản, phản ánh rõ đặc điểm của các biểu cảm khuôn mặt trong tập dữ liệu.



Hình 8: Trực quan hóa dữ liệu FER-2013 bằng PCA.

Hình 8 thể hiện việc trực quan hóa tập dữ liệu FER-2013 bằng phương pháp phân tích thành phần chính (PCA). Hai thành phần chính đầu tiên được chọn để biểu diễn dữ liệu trong không gian 2 chiều, giúp giảm chiều và dễ dàng quan sát hơn. Mỗi điểm trong biểu đồ đại diện cho một ảnh, được tô màu theo nhãn cảm xúc (từ 0 đến 6). Có thể nhận thấy rằng các điểm dữ liệu phân bố khá chồng chéo nhau, cho thấy các lớp cảm xúc trong FER-2013 có sự giao thoa đáng kể và không tách biệt rõ ràng trong không gian PCA, điều này phản ánh tính chất phức tạp và khó phân biệt giữa các biểu cảm khuôn mặt.

3.3.2 Tiền xử lý dữ liệu

Các bước tiền xử lý được thực hiện nhằm cải thiện chất lượng ảnh đầu vào và mô phỏng các điều kiện môi trường khác nhau, cụ thể như sau:

- Chuẩn hóa hình ảnh: Loại bỏ nhiễu và đảm bảo định dạng đồng nhất (kích thước ảnh, không gian màu), giúp mô hình huấn luyện ổn định hơn.
- Mô phỏng điều kiện ánh sáng yếu: Để mô phỏng môi trường có ánh sáng yếu, hình ảnh được chuyển sang không gian màu HSV và kênh độ sáng (V) được

giảm xuống còn 10% so với ảnh gốc. Sau đó, ảnh được chuyển lại về không gian RGB để sử dụng trong huấn luyện.

3.3.3 Tăng cường dữ liệu thích ứng

Áp dụng các phép biến đổi linh hoạt dựa trên đặc trưng ánh sáng của từng ảnh. Việc tăng cường được thực hiện bằng Python với OpenCV và NumPy.

3.4 Phân tích dữ liệu

3.4.1 Công cụ và phần mềm

- Python: xử lý dữ liệu và huấn luyện mô hình.
- PyTorch : xử lý dữ liệu và huấn luyện mô hình.
- TensorFlow/Keras: xây dựng và đánh giá mô hình.
- OpenCV: tiền xử lý ảnh.
- NumPy, Pandas: quản lý dữ liệu.
- Matplotlib, Seaborn: trực quan hóa kết quả.

3.4.2 Quy trình phân tích

- Huấn luyện mô hình MobileNetV3Small:
  - Sử dụng mô hình MobileNetV3Small với trọng số ImageNet, loại bỏ phần fully-connected gốc (include_top=False).
  - Chỉ tinh chỉnh 30 lớp cuối cùng trong mạng, các lớp còn lại được đóng băng để giữ lại đặc trưng đã học.
  - Kiến trúc phần đầu ra gồm: Global Average Pooling, hai lớp Dense (128 và 64 nodes, activation ReLU), kèm Dropout 0.3, kết thúc bằng lớp Softmax với 7 nhãn đầu ra.
  - Hàm mất mát: Categorical Crossentropy.
  - Tối ưu hóa bằng Adam (learning rate mặc định).
  - Số epoch: 10, sử dụng Early Stopping với patience = 3 để tránh overfitting.
- Huấn luyện mô hình với ResNet18:

  - Sử dụng mô hình ResNet18 với trọng số đã được huấn luyện sẵn trên tập ImageNet (ResNet18_Weights.IMAGENET1K_V1).
  - Điều chỉnh lại lớp Fully Connected cuối cùng thành nn.Linear(..., 7) để phù hợp với bài toán phân loại 7 cảm xúc trên tập dữ liệu FER2013.
  - Hàm mất mát sử dụng là CrossEntropyLoss, phù hợp với phân loại đa lớp.
  - Trình tối ưu hóa: Adam với learning rate 0.001.
  - Mô hình được huấn luyện trong 20 epoch.
  - Trong quá trình huấn luyện, độ chính xác và mất mát (loss) trên tập huấn luyện và tập xác thực được theo dõi để đánh giá hiệu quả mô hình. Mô hình tốt nhất được lưu lại sau mỗi epoch nếu có cải thiện.
- Đánh giá mô hình:
  - Các chỉ số đánh giá: Accuracy, Precision, Recall, F1-score.
  - Đo thời gian suy luận trung bình trên CPU (per image).
  - Kích cỡ mô hình sau huấn luyện.
- So sánh mô hình:
  - MobileNetV3 (cơ bản vs. tăng cường).
  - ResNet18 (cơ bản vs. tăng cường).
  - So sánh giữa MobileNetV3 và ResNet18.
- Phân tích kết quả:
  - Ma trận nhầm lẫn, biểu đồ Accuracy theo epoch.
  - Quan sát các trường hợp dự đoán sai.

3.4.3 Thiết bị triển khai

Thực nghiệm được thực hiện trên máy MacBook Air M1, được trang bị chip Apple M1 và RAM 8GB. Ngoài ra, Google Colab cũng được sử dụng để mô phỏng điều kiện tài nguyên thấp, với việc chỉ sử dụng CPU thay vì GPU nhằm đánh giá thời gian suy luận, phù hợp với môi trường nhúng.

3.5 Phương pháp so sánh

Nghiên cứu tiến hành so sánh định lượng qua các chỉ số hiệu suất (Accuracy, Precision, Recall, F1-score) và thời gian suy luận giữa:

- MobileNetV3 cơ bản vs. tăng cường.
- ResNet18 cơ bản vs. tăng cường.
- So sánh giữa MobileNetV3 và ResNet18.

Kết quả được trình bày dưới dạng bảng và biểu đồ để làm rõ hiệu quả của các kỹ thuật và sự phù hợp của mô hình trong ứng dụng thực tế.

4 Thực nghiệm và thảo luận

4.1 Thiết lập thực nghiệm

Trong phần này, chúng tôi tiến hành đánh giá hiệu suất của hai mô hình học sâu là MobileNetV3Small và ResNet18 trong bài toán phân loại cảm xúc khuôn mặt trên tập dữ liệu FER2013. Mỗi mô hình được thử nghiệm trên ba biến thể của tập dữ liệu nhằm khảo sát khả năng thích nghi với điều kiện ánh sáng thay đổi và hiệu quả của các kỹ thuật tăng cường dữ liệu.

Bảng 2: Kiến trúc mô hình MobileNetV3Small sử dụng trong thực nghiệm

Layer (type)	Output Shape	Param #
MobileNetV3Small (Functional)	(None, 7, 7, 576)	939,120
GlobalAveragePooling2D	(None, 576)	0
Dense (128 units)	(None, 128)	73,856
Dropout	(None, 128)	0
Dense (64 units)	(None, 64)	8,256
Dropout	(None, 64)	0
Dense (7 units - output)	(None, 7)	455

Bảng 2 mô tả kiến trúc của mô hình MobileNetV3Small được sử dụng trong thực nghiệm. Mô hình gốc MobileNetV3Small được sử dụng như một bộ trích xuất đặc trưng (feature extractor) đầu vào với đầu ra có kích thước (7, 7, 576). Sau đó, lớp GlobalAveragePooling2D được áp dụng để giảm chiều không gian, tạo vector đặc trưng một chiều với 576 phần tử.

Tiếp theo là hai lớp Dense với số lượng đơn vị lần lượt là 128 và 64, đi kèm với các lớp Dropout nhằm giảm hiện tượng overfitting. Cuối cùng, lớp Dense đầu ra có 7 đơn vị tương ứng với 7 lớp cảm xúc cần phân loại.

Tổng số tham số huấn luyện của toàn bộ mô hình là khoảng 1 triệu, trong đó phần lớn nằm ở MobileNetV3Small. Việc sử dụng kiến trúc gọn nhẹ giúp mô hình đạt được hiệu quả cao mà vẫn đảm bảo tốc độ xử lý nhanh, phù hợp với các ứng dụng thực tế như trên thiết bị di động.

Bảng 3: Cấu trúc mô hình mạng học sâu

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 64, 112, 112]	9,408
BatchNorm2d-2	[-1, 64, 112, 112]	128
ReLU-3	[-1, 64, 112, 112]	0
MaxPool2d-4	[-1, 64, 56, 56]	0
Conv2d-5	[-1, 64, 56, 56]	36,864
BatchNorm2d-6	[-1, 64, 56, 56]	128
ReLU-7	[-1, 64, 56, 56]	0
Conv2d-8	[-1, 64, 56, 56]	36,864
BatchNorm2d-9	[-1, 64, 56, 56]	128
ReLU-10	[-1, 64, 56, 56]	0
BasicBlock-11	[-1, 64, 56, 56]	0
Conv2d-12	[-1, 64, 56, 56]	36,864
BatchNorm2d-13	[-1, 64, 56, 56]	128
ReLU-14	[-1, 64, 56, 56]	0
Conv2d-15	[-1, 64, 56, 56]	36,864
BatchNorm2d-16	[-1, 64, 56, 56]	128
ReLU-17	[-1, 64, 56, 56]	0
BasicBlock-18	[-1, 64, 56, 56]	0
Conv2d-19	[-1, 128, 28, 28]	73,728
BatchNorm2d-20	[-1, 128, 28, 28]	256

(tiếp trang sau)

Layer (type)	Output Shape	Param #
ReLU-21	[-1, 128, 28, 28]	0
Conv2d-22	[-1, 128, 28, 28]	147,456
BatchNorm2d-23	[-1, 128, 28, 28]	256
Conv2d-24	[-1, 128, 28, 28]	8,192
BatchNorm2d-25	[-1, 128, 28, 28]	256
ReLU-26	[-1, 128, 28, 28]	0
BasicBlock-27	[-1, 128, 28, 28]	0
Conv2d-28	[-1, 128, 28, 28]	147,456
BatchNorm2d-29	[-1, 128, 28, 28]	256
ReLU-30	[-1, 128, 28, 28]	0
Conv2d-31	[-1, 128, 28, 28]	147,456
BatchNorm2d-32	[-1, 128, 28, 28]	256
ReLU-33	[-1, 128, 28, 28]	0
BasicBlock-34	[-1, 128, 28, 28]	0
Conv2d-35	[-1, 256, 14, 14]	294,912
BatchNorm2d-36	[-1, 256, 14, 14]	512
ReLU-37	[-1, 256, 14, 14]	0
Conv2d-38	[-1, 256, 14, 14]	589,824
BatchNorm2d-39	[-1, 256, 14, 14]	512
Conv2d-40	[-1, 256, 14, 14]	32,768
BatchNorm2d-41	[-1, 256, 14, 14]	512
ReLU-42	[-1, 256, 14, 14]	0
BasicBlock-43	[-1, 256, 14, 14]	0
Conv2d-44	[-1, 256, 14, 14]	589,824
BatchNorm2d-45	[-1, 256, 14, 14]	512
ReLU-46	[-1, 256, 14, 14]	0
Conv2d-47	[-1, 256, 14, 14]	589,824
BatchNorm2d-48	[-1, 256, 14, 14]	512
ReLU-49	[-1, 256, 14, 14]	0

(tiếp trang sau)

Layer (type)	Output Shape	Param #
BasicBlock-50	[-1, 256, 14, 14]	0
Conv2d-51	[-1, 512, 7, 7]	1,179,648
BatchNorm2d-52	[-1, 512, 7, 7]	1,024
ReLU-53	[-1, 512, 7, 7]	0
Conv2d-54	[-1, 512, 7, 7]	2,359,296
BatchNorm2d-55	[-1, 512, 7, 7]	1,024
Conv2d-56	[-1, 512, 7, 7]	131,072
BatchNorm2d-57	[-1, 512, 7, 7]	1,024
ReLU-58	[-1, 512, 7, 7]	0
BasicBlock-59	[-1, 512, 7, 7]	0
Conv2d-60	[-1, 512, 7, 7]	2,359,296
BatchNorm2d-61	[-1, 512, 7, 7]	1,024
ReLU-62	[-1, 512, 7, 7]	0
Conv2d-63	[-1, 512, 7, 7]	2,359,296
BatchNorm2d-64	[-1, 512, 7, 7]	1,024
ReLU-65	[-1, 512, 7, 7]	0
BasicBlock-66	[-1, 512, 7, 7]	0
AdaptiveAvgPool2d-67	[-1, 512, 1, 1]	0
Linear-68	[-1, 7]	3,591
Total Params		11,180,103
Trainable Params		11,180,103
Non-trainable Params		0

Mô hình sử dụng là một biến thể của kiến trúc ResNet18 với tổng cộng 68 tầng. Các tầng chính trong mô hình bao gồm:

- Conv2d: Tầng tích chập, giúp trích xuất đặc trưng từ ảnh đầu vào.
- BatchNorm2d: Chuẩn hóa các giá trị đầu ra theo batch, giúp mô hình hội tụ nhanh hơn.

- ReLU: Hàm kích hoạt phi tuyến tính, tăng khả năng biểu diễn của mô hình.
- MaxPool2d: Gộp cực đại, giảm kích thước không gian và giữ lại đặc trưng quan trọng.
- BasicBlock: Khối residual trong ResNet giúp truyền gradient hiệu quả, giảm hiện tượng mất mát gradient trong mạng sâu.
- AdaptiveAvgPool2d: Lớp pooling trung bình thích ứng, đưa kích thước về dạng cố định để chuẩn bị cho tầng fully connected.
- Linear: Lớp kết nối đầy đủ (fully connected) để thực hiện phân loại đầu ra.

Lớp đầu ra (Output layer) của mô hình là Linear-68 với 7 đơn vị đầu ra, tương ứng với 7 lớp cảm xúc trong bài toán phân loại (ví dụ: Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise).

Khối BasicBlock: ResNet18 sử dụng các khối residual (BasicBlock) để khắc phục hiện tượng suy giảm gradient trong các mạng học sâu. Trong mô hình này, mỗi giai đoạn chứa 2 BasicBlock, ngoại trừ giai đoạn đầu tiên. Tổng cộng mô hình sử dụng 16 khối BasicBlock, phù hợp với kiến trúc gốc của ResNet18.

Tổng số tham số huấn luyện của toàn bộ mô hình là khoảng 11,180,103 tham số

4.2 Biểu đồ, bảng biểu, hình ảnh minh hoạ

4.2.1 MobileNetV3Small

Bảng 4: Độ chính xác của các phiên bản mô hình MobileNetV3Small trên tập dữ liệu FER2013

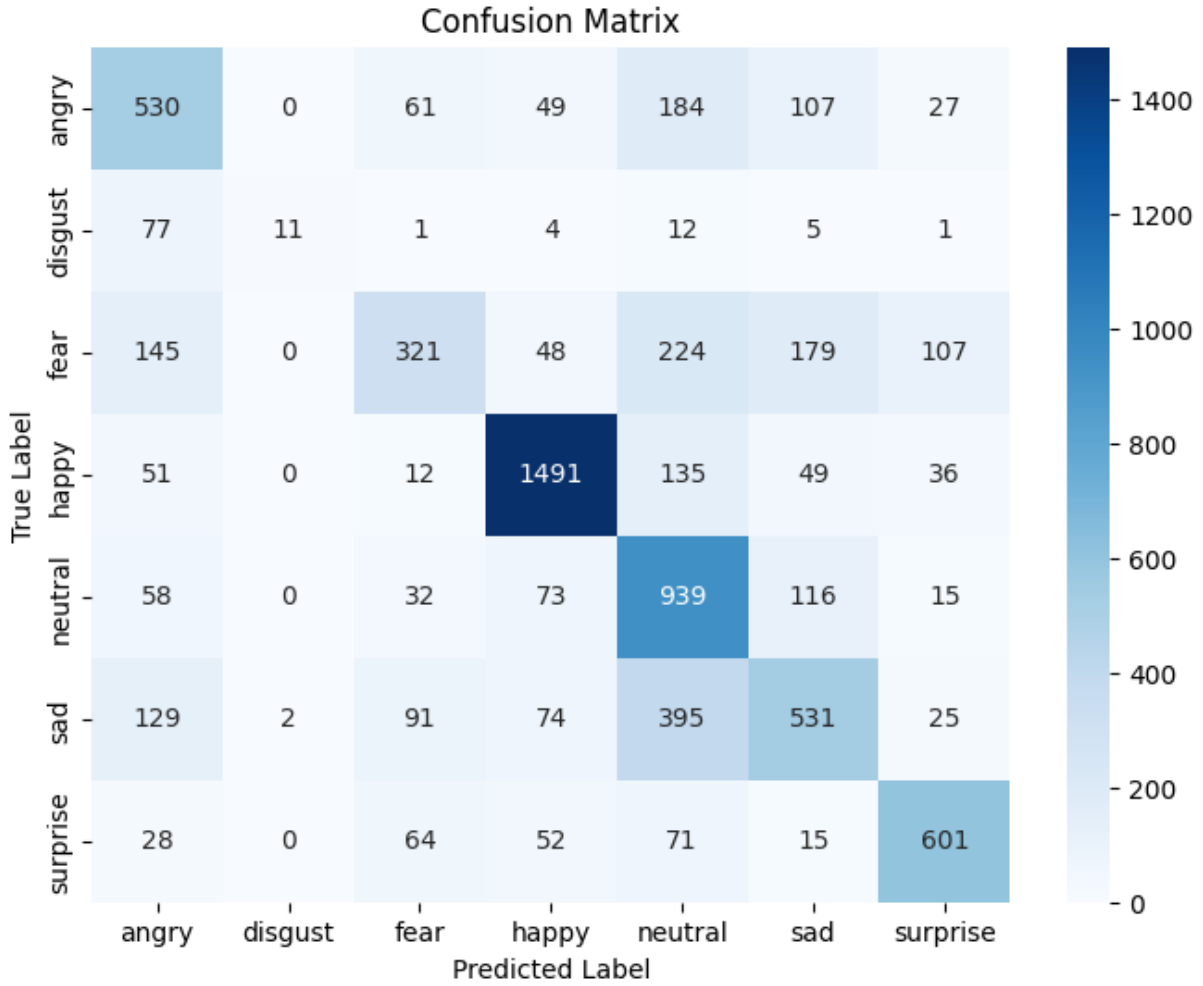
Tên kiến trúc	Độ chính xác (%)
MobileNetV3Small + FER2013	61.63
MobileNetV3Small + FER2013 (Low Light Images - LLI)	58.86
MobileNetV3Small + FER2013 (LLI + adaptive augmentation)	61.55

Bảng 4 trình bày độ chính xác khi huấn luyện mô hình MobileNetV3Small trên các phiên bản khác nhau của tập dữ liệu FER2013. Trong đó, FER2013 là tập dữ liệu gốc chứa các hình ảnh khuôn mặt thể hiện cảm xúc.

Khi huấn luyện trên tập FER2013 gốc, mô hình đạt độ chính xác cao nhất

là 61.63%. Việc mô phỏng điều kiện ánh sáng yếu thông qua tập dữ liệu LLI làm giảm độ chính xác mô hình, phản ánh thách thức trong việc nhận diện cảm xúc dưới điều kiện chiếu sáng kém, giảm còn 58.86%. Tuy nhiên, khi kết hợp LLI với phương pháp tăng cường dữ liệu thích ứng (adaptive augmentation), độ chính xác cải thiện đáng kể lên 61.55%, gần tương đương với mô hình gốc.

Điều này cho thấy rằng các kỹ thuật tăng cường dữ liệu phù hợp có thể giúp mô hình thích nghi tốt hơn với điều kiện ánh sáng kém, nhưng cần được áp dụng và điều chỉnh một cách hợp lý để tránh làm nhiễu thông tin đầu vào.



Hình 9: Ma trận nhầm lẫn trên tập kiểm tra của mô hình MobileNetV3

Hình 9 trình bày ma trận nhầm lẫn của mô hình MobileNetV3 trong nhiệm vụ phân loại cảm xúc khuôn mặt. Kết quả cho thấy mô hình nhận diện tốt các cảm xúc có đặc trưng rõ ràng như:

- Happy: có 1,491 mẫu được dự đoán đúng (95.40%).
- Neutral: có 939 mẫu đúng (64.22%).
- Surprise: đạt 601 mẫu đúng (86.48%).

Tuy nhiên, hiệu suất giảm đáng kể với các cảm xúc khó phân biệt hơn:

- Disgust: chỉ 11 mẫu được nhận diện đúng (21.57%), trong khi bị nhầm sang Angry tới 28 mẫu (54.90%).
- Fear: chỉ 42 mẫu đúng (8.57%), trong khi bị nhầm với:
  - Neutral: 224 mẫu (45.71%)
  - Sad: 179 mẫu (36.53%)
  - Angry: 145 mẫu (29.59%)
- Sad: chỉ 107 mẫu đúng (17.63%), bị nhầm với:
  - Neutral: 395 mẫu (65.07%)
  - Fear: 91 mẫu (14.98%)

Tổng thể, ma trận nhầm lẫn cung cấp cái nhìn rõ nét về khả năng mô hình phân biệt giữa các cảm xúc. Trong khi các cảm xúc tích cực như happy và surprise đạt hiệu suất cao, các cảm xúc tiêu cực như fear, sad và disgust dễ bị nhầm lẫn lẫn nhau, đòi hỏi cải tiến thêm về dữ liệu huấn luyện và biểu diễn đặc trưng.

4.2.2 ResNet18

Bảng 5: Độ chính xác của các phiên bản mô hình ResNet18 trên tập dữ liệu FER2013

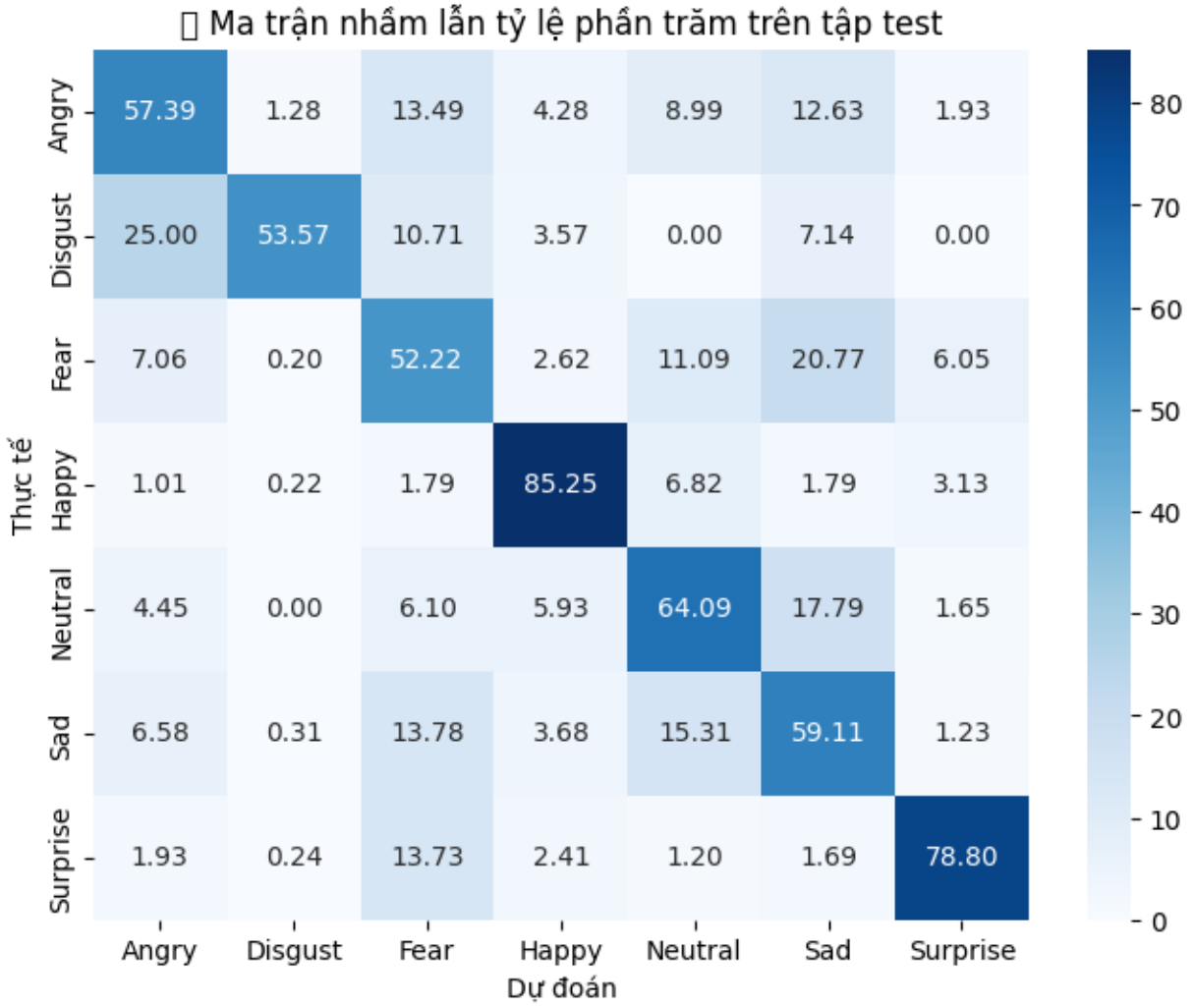
Tên kiến trúc	Độ chính xác (%)
ResNet18 + FER2013	67.23
ResNet18 + FER2013 (Low Light Images - LLI)	67.04
ResNet18 + FER2013 (LLI + adaptive augmentation)	67.48

Bảng 5 thể hiện độ chính xác của mô hình ResNet18 khi huấn luyện trên các phiên bản khác nhau của tập dữ liệu FER2013. Tương tự như mô hình MobileNetV3Small, FER2013 là tập dữ liệu ảnh khuôn mặt thể hiện cảm xúc.

Mô hình ResNet18 huấn luyện trên tập FER2013 gốc đạt độ chính xác cao

nhất là 67.23%. Khi sử dụng phiên bản ảnh ánh sáng yếu (Low Light Images - LLI), độ chính xác chỉ giảm nhẹ còn 67.04%. Đáng chú ý, việc kết hợp thêm phương pháp tăng cường dữ liệu thích ứng (adaptive augmentation) giúp cải thiện hiệu suất lên mức 67.48%, vượt cả mô hình gốc.

Kết quả này cho thấy ResNet18 có khả năng học tốt trong điều kiện ánh sáng kém và có thể tận dụng tốt lợi ích từ các kỹ thuật tăng cường dữ liệu phù hợp.



Hình 10: Ma trận nhầm lẫn phần trăm trên tập kiểm tra của mô hình ResNet18

Phân tích ma trận nhầm lẫn thể hiện tại Hình 10 cho thấy hiệu quả phân loại cảm xúc của mô hình ResNet18 trên tập kiểm tra. Mô hình đạt độ chính xác cao đối với các cảm xúc như Happy (85.25%) và Surprise (78.80%), tương đồng với các nghiên cứu trước đó về độ dễ nhận diện của các cảm xúc này qua biểu cảm

khuôn mặt.

Ngược lại, các cảm xúc như Fear, Sad và Disgust có xu hướng bị nhầm lẫn nhiều, đặc biệt là giữa các cặp có biểu hiện tương tự về hình thái học (ví dụ: Fear và Sad). Điều này cho thấy mô hình vẫn còn hạn chế trong việc tách biệt các đặc trưng tinh vi giữa các cảm xúc gần nhau về biểu cảm, một thách thức phổ biến trong nhận diện cảm xúc.

Những kết quả này gợi ý rằng việc cải thiện mô hình có thể tập trung vào việc xử lý mất cân bằng lớp, sử dụng dữ liệu tăng cường phù hợp và áp dụng kỹ thuật học sâu nâng cao để tăng cường khả năng phân biệt giữa các cảm xúc dễ gây nhầm lẫn.

4.3 Đánh giá, giải thích kết quả nghiên cứu

Bảng 6: Kết quả tổng thể của các mô hình

Tên kiến trúc	Accuracy (%)	F1-score (%)	Time infer (ms)	Size (MB)
MobileNetV3Small	61.63	60	2.15	13.54
ResNet18	67.23	67	3.44	42.73
MobileNetV3Small + FER2013 LLI	58.86	58	2.10	13.54
ResNet18 + FER2013 LLI	67.04	67	3.18	42.72
MobileNetV3Small + FER2013 LLI + adaptive augmentation	61.55	60	2.12	13.54
ResNet18 + FER2013 LLI + adaptive augmentation	67.48	67	2.91	42.72

Bảng trên tổng hợp các chỉ số chính của các mô hình thử nghiệm, bao gồm độ chính xác (Accuracy), F1-score, thời gian suy luận trên mỗi mẫu (Time inference), và kích thước mô hình (Size).

Nhận xét:

- Hiệu suất nhận diện: Mô hình ResNet18 cho kết quả vượt trội hơn hẳn MobileNetV3Small về độ chính xác và F1-score, chứng tỏ khả năng biểu diễn đặc trưng sâu sắc và hiệu quả hơn cho bài toán phân loại cảm xúc.
- Ảnh hưởng của dữ liệu ánh sáng yếu (LLI): Cả hai mô hình đều giảm nhẹ độ chính xác khi huấn luyện trên tập dữ liệu LLI, tuy nhiên sự sụt giảm này

không đáng kể, đặc biệt là với ResNet18. Điều này cho thấy các kiến trúc mạng này có độ bền tốt với các biến đổi về điều kiện ánh sáng.

- Tác động của kỹ thuật tăng cường dữ liệu thích ứng: Khi kết hợp LLI với adaptive augmentation, độ chính xác của cả hai mô hình đều được cải thiện và gần đạt bằng hoặc vượt mức mô hình gốc. Đặc biệt với ResNet18, mức cải thiện rõ ràng và thời gian suy luận cũng giảm nhẹ, chứng tỏ kỹ thuật tăng cường dữ liệu giúp mô hình học được đặc trưng phong phú hơn và tăng khả năng tổng quát hóa.
- Hiệu năng và kích thước mô hình: MobileNetV3Small có lợi thế vượt trội về kích thước nhỏ gọn và thời gian suy luận nhanh hơn, phù hợp cho các ứng dụng thực tế trên thiết bị di động hoặc các hệ thống giới hạn tài nguyên. Ngược lại, ResNet18 với hiệu suất cao hơn đi kèm kích thước và thời gian suy luận lớn hơn, thích hợp cho các hệ thống yêu cầu độ chính xác cao.

Kết luận:

Việc lựa chọn mô hình phụ thuộc vào mục tiêu ứng dụng cụ thể. Nếu ưu tiên độ chính xác và khả năng xử lý đa dạng điều kiện ánh sáng, ResNet18 là lựa chọn tối ưu. Trong khi đó, MobileNetV3Small phù hợp cho các ứng dụng cần tối ưu tài nguyên và tốc độ xử lý. Bên cạnh đó, áp dụng kỹ thuật tăng cường dữ liệu thích ứng được khuyến khích để cải thiện độ bền mô hình trong các môi trường ánh sáng thay đổi, giúp tăng khả năng nhận diện cảm xúc chính xác và ổn định hơn.

4.4 Nêu ý nghĩa thực tiễn của nghiên cứu

- Cải thiện độ chính xác trong điều kiện ánh sáng yếu: Hệ thống nhận diện cảm xúc thường gặp khó khăn khi môi trường thiếu sáng (ví dụ: buổi tối, trong xe hơi, phòng họp mờ...). Nghiên cứu này giúp khắc phục vấn đề đó, tăng độ tin cậy và ổn định của mô hình trong điều kiện thực tế.
- Giảm chi phí phần cứng: Thay vì cần máy ảnh chất lượng cao để chụp rõ trong điều kiện ánh sáng kém, việc tăng cường ảnh bằng phần mềm cho phép sử dụng thiết bị giá rẻ, phù hợp với triển khai diện rộng.

5 Kết luận và hướng phát triển

5.1 Tóm tắt kết quả nghiên cứu

Nghiên cứu đã xây dựng thành công một hệ thống nhận diện biểu cảm khuôn mặt trong điều kiện ánh sáng yếu, dựa trên mô hình CNN nhẹ MobileNetV3 kết hợp với pipeline tăng cường dữ liệu thích ứng. Phương pháp xử lý ảnh thích ứng dựa trên phân tích đặc trưng độ sáng của ảnh đầu vào đã góp phần nâng cao chất lượng hình ảnh, từ đó cải thiện đáng kể hiệu suất nhận diện. Kết quả thực nghiệm trên tập dữ liệu FER-2013 cho thấy mô hình đạt độ chính xác 61.55

Nghiên cứu cũng chỉ ra một số hạn chế, bao gồm độ chính xác tổng thể còn khiêm tốn và hiện tượng nhầm lẫn giữa các biểu cảm có đặc trưng tương đồng, nguyên nhân chủ yếu do mất cân bằng dữ liệu và sự gần giống của các biểu cảm.

5.2 Câu hỏi nghiên cứu

Nghiên cứu tập trung trả lời các câu hỏi sau:

1. Làm thế nào để thiết kế một pipeline tăng cường dữ liệu thích ứng, hiệu quả trong việc cải thiện chất lượng ảnh ánh sáng yếu phục vụ nhận diện biểu cảm khuôn mặt?
2. Mô hình MobileNetV3 có thể đạt được độ chính xác tương đương hoặc vượt trội so với các kỹ thuật tăng cường dữ liệu cố định trong điều kiện ánh sáng yếu hay không?
3. Các kỹ thuật tăng cường dữ liệu thích ứng ảnh hưởng như thế nào đến hiệu suất của mô hình CNN nhẹ trong nhiệm vụ nhận diện biểu cảm khuôn mặt?

Thông qua quá trình nghiên cứu và thử nghiệm, các câu hỏi trên được giải đáp như sau:

- Pipeline tăng cường dữ liệu thích ứng được thiết kế dựa trên phân tích độ sáng trung bình và độ lệch chuẩn của ảnh đầu vào, giúp điều chỉnh linh hoạt các phép biến đổi ảnh phù hợp với đặc điểm từng ảnh. Giải pháp này đã chứng minh hiệu quả trong việc cải thiện chất lượng ảnh ánh sáng yếu, từ đó nâng cao độ chính xác nhận diện.

- Mô hình MobileNetV3 khi kết hợp với kỹ thuật tăng cường dữ liệu thích ứng đạt hiệu suất gần bằng với mô hình trên ảnh gốc và vượt trội so với kỹ thuật tăng cường dữ liệu cố định, chứng tỏ khả năng thích nghi tốt với các điều kiện ánh sáng khác nhau.
- Kỹ thuật tăng cường dữ liệu thích ứng giúp giảm thiểu hiện tượng mất chi tiết trên khuôn mặt trong môi trường thiếu sáng, từ đó cải thiện hiệu suất tổng thể của mô hình CNN nhẹ, mặc dù vẫn tồn tại một số hạn chế liên quan đến sự nhầm lẫn giữa các biểu cảm có đặc trưng tương đồng.

5.3 Hướng phát triển

Trên cơ sở kết quả đạt được và các hạn chế còn tồn tại, nghiên cứu đề xuất các hướng phát triển tiếp theo như sau:

- Mở rộng tập dữ liệu ánh sáng yếu đa dạng hơn về biểu cảm, độ tuổi, giới tính và điều kiện môi trường nhằm nâng cao khả năng tổng quát hóa của mô hình.
- Nghiên cứu và áp dụng các kỹ thuật tiền xử lý ảnh hiệu quả hơn, đảm bảo tốc độ xử lý nhanh và phù hợp với thiết bị nhúng.
- Tối ưu hóa mô hình CNN nhẹ thông qua các phương pháp lượng tử hóa, cắt tỉa hoặc kiến trúc mới nhằm giảm kích thước và thời gian suy luận, đồng thời duy trì hoặc nâng cao độ chính xác.
- Phát triển các kỹ thuật học đặc trưng nâng cao hoặc mô hình đa nhiệm để cải thiện khả năng phân biệt các biểu cảm tương đồng.
- Triển khai và đánh giá hệ thống trên các thiết bị thực tế trong môi trường ánh sáng yếu nhằm xác định tính khả thi và hiệu quả ứng dụng thực tiễn.

Nghiên cứu kỳ vọng sẽ góp phần thúc đẩy phát triển các giải pháp nhận diện biểu cảm khuôn mặt hiệu quả và khả thi trong điều kiện ánh sáng yếu, phục vụ các ứng dụng đa dạng trong lĩnh vực giám sát an ninh, tương tác người - máy và chăm sóc sức khỏe tâm lý.

TÀI LIỆU THAM KHẢO

[1] S. Kusal et al., "A review on text-based emotion detection—techniques, applications, datasets, and future directions," arXiv preprint, arXiv:2205.03235, 2022.

[2] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "URetinex-Net: Retinex-based deep unfolding network for low-light image enhancement," in Proc. IEEE CVPR, 2022, pp. 5901–5910.

[3] M. Bie et al., "DA-FER: Domain adaptive facial expression recognition," Appl. Sci., vol. 13, no. 10, p. 6314, 2023, doi: 10.3390/app13106314.

[4] L. A. Al Hak, W. A. Ali, and S. J. Saba, "Facial expression recognition using data augmentation and transfer learning," Ingénierie des Systèmes d'Information, vol. 29, no. 3, pp. 1219–1225, 2024, doi: 10.18280/isi.290338.

[5] A. G. Howard et al., "Searching for MobileNetV3," in Proc. IEEE ICCV, 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.

[6] X. Liang, J. Liang, T. Yin, and X. Tang, "A lightweight method for face expression recognition based on improved MobileNetV3," IET Image Process., vol. 17, no. 8, pp. 2375–2384, 2023, doi: 10.1049/ipe2.12798.

[7] S. B. R. Prasad and B. S. Chandana, "MobileNetV3: A deep learning technique for human face expressions identification," Int. J. Inf. Technol., 2023, doi: 10.1007/s41870-023-01380-x.

[8] Z. Zhang et al., "EnlightenGAN: Deep light enhancement GAN for low-light images," ACM MM, 2019.

[9] L. Ying et al., "A bio-inspired multi-exposure fusion framework for low-light image enhancement," IEEE TIP, 2017.

[10] Y. Chen et al., "An adaptive preprocessing method for low-light FER," Pattern Recognition Letters, vol. 150, pp. 92–99, 2021.

[11] J. Hu et al., "Squeeze-and-excitation networks," IEEE CVPR, 2018.

[12] W. Zhou et al., "Adaptive data augmentation for facial expression recognition," Neurocomputing, vol. 453, 2021.