APPLICATION OF GENERATIVE ADVERSARIAL NETWORKS (GANS) TO ENHANCE FACIAL EMOTION CLASSIFICATION PERFORMANCE USING CONVOLUTIONAL NEURAL NETWORKS (CNNS)

ỨNG DỤNG MẠNG ĐỐI SINH (GAN) ĐỂ NÂNG CAO HIỆU SUẤT PHÂN LOẠI CẢM XÚC KHUÔN MẶT NGƯỜI BẰNG MẠNG NƠ-RON TÍCH CHẬP (CNN)

Nguyễn Thị Hà Phương, Nguyễn Thị Hải Lê
Khoa Kỹ thuật và Công nghệ, Đại học Huế

ABSTRACT: *Currently, human face recognition is a widely studied research area and has achieved significant advancements. Among various approaches, modern deep learning techniques such as Convolutional Neural Networks (CNNs) have demonstrated relatively high recognition performance. However, the effectiveness of deep learning models like CNNs heavily depends on the quality and quantity of training data - especially in tasks such as* ***facial emotion classification****, where datasets are often imbalanced or limited. To address this limitation, we propose using* ***Generative Adversarial Networks (GANs)*** *to synthesize artificial facial images with diverse and realistic emotional expressions. These synthetic images are then utilized to augment the training dataset for a CNN-based facial emotion classification model, specifically using the VGG-16 architecture. Integrating of GAN-generated data enhances the model's generalization capability and improves classification accuracy. Experimental results show that the VGG-16 model trained with GAN-augmented data achieves an accuracy of* ***95.3%****, which is* ***3.6% higher*** *than the model trained solely on the original dataset (****91.7%****). This confirms the effectiveness of applying GANs to the facial emotion classification task, outperforming traditional data augmentation methods and contributing to the improved performance of human face recognition systems.*

Keywords: *Convolutional Neural Network-CNN, Generative adversarial Network - GAN, VGG-16, Data Augmentation, Facial Emotion Classification.*

TÓM TẮT: *Hiện nay, nhận dạng khuôn mặt người là một lĩnh vực nghiên cứu phổ biến và đã đạt được nhiều tiến bộ. Trong đó, việc nhận dạng bằng phương pháp học sâu như mạng nơ-ron tích chập (CNN) đem lại hiệu suất nhận dạng tương đối cao. Tuy nhiên, hiệu quả của các mô hình học sâu như CNN phụ thuộc mạnh mẽ vào chất lượng và số lượng dữ liệu huấn luyện, đặc biệt trong các bài toán như* ***phân loại cảm xúc khuôn mặt****, nơi dữ liệu thường mất cân bằng hoặc hạn chế. Để khắc phục hạn chế này, chúng tôi đề xuất sử dụng mạng đối sinh (GAN) để sinh ra các hình ảnh khuôn mặt giả có đặc trưng cảm xúc đa dạng và tương đối chân thực. Các hình ảnh tổng hợp này sau đó được sử dụng để* ***tăng cường tập huấn luyện*** *cho mô hình phân loại cảm xúc khuôn mặt dựa trên CNN, cụ thể là kiến trúc VGG-16. Việc tích hợp dữ liệu sinh bởi GAN giúp cải thiện khả năng khái quát hoá của mô hình và nâng cao độ chính xác phân loại cảm xúc. Kết quả thực nghiệm cho thấy, mô hình VGG-16 khi được huấn luyện với dữ liệu tăng cường từ GAN đạt độ chính xác* ***95,3%****, cao hơn* ***3,6%*** *so với mô hình chỉ huấn luyện với dữ liệu gốc (****91,7%****). Điều này khẳng định hiệu quả của việc ứng dụng GAN trong bài toán phân loại cảm xúc khuôn mặt, từ đó góp phần nâng cao hiệu suất các hệ thống nhận dạng khuôn mặt người.*

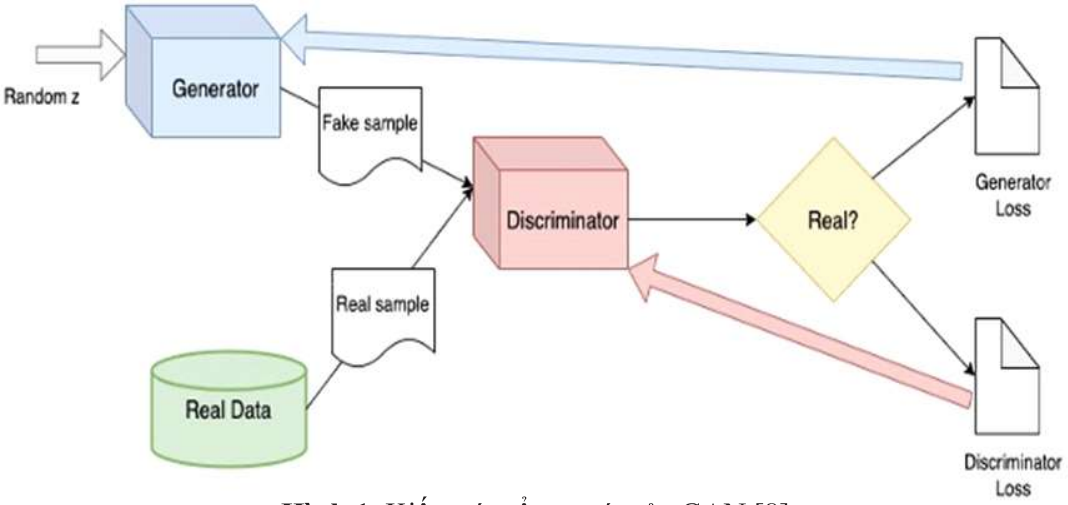
Từ khóa: *Convolutional Neural Network-CNN, Generative adversarial Network - GAN, VGG-16, Tăng cường dữ liệu, Phân loại cảm xúc khuôn mặt.*

1. ĐẶT VẤN ĐỀ

Nhận dạng khuôn mặt là một lĩnh vực quan trọng trong thị giác máy tính, với nhiều ứng dụng thực tiễn như bảo mật, giám sát, tương tác người - máy và xác thực danh tính. Trong những năm gần đây, các phương pháp học sâu, đặc biệt là mạng nơ-ron tích chập (CNN), đã đạt được những thành tựu đáng kể trong việc cải thiện độ chính xác của các hệ thống nhận dạng khuôn mặt nhờ khả năng học đặc trưng tự động từ dữ liệu huấn luyện lớn [15] [10] [16].

Tuy nhiên, hiệu suất của các mô hình học sâu như CNN phụ thuộc mạnh mẽ vào chất lượng và số lượng dữ liệu huấn luyện. Trong các bài toán như phân loại cảm xúc khuôn mặt, dữ liệu thường bị mất cân bằng hoặc hạn chế, dẫn đến hiện tượng overfitting và giảm khả năng khái quát hóa của mô hình [14], [15]. Việc thu thập dữ liệu thực tế với đầy đủ các biến thể về góc nhìn, điều kiện ánh sáng, biểu cảm và che khuất không chỉ tốn kém mà còn gặp phải nhiều rào cản về quyền riêng tư.

Để khắc phục hạn chế này, các phương pháp tăng cường dữ liệu (Data Augmentation) đã được áp dụng rộng rãi. Các kỹ thuật truyền thống như lật ảnh, xoay, thay đổi độ sáng, thêm nhiễu và cắt xén đã chứng minh hiệu quả trong việc tăng tính đa dạng của dữ liệu huấn luyện và cải thiện hiệu suất của mô hình [11]. Một khảo sát toàn diện về các phương pháp tăng cường dữ liệu hiện đại đã được trình bày trong nghiên cứu [14]. Tuy nhiên, các phương pháp tăng cường truyền thống có thể không đủ để tạo ra các biến thể phức tạp và chân thực của khuôn mặt. Trong bối cảnh này, mạng đối sinh (GAN - Generative Adversarial Networks) đã nổi lên như một công cụ mạnh mẽ để tạo ra dữ liệu tổng hợp có tính chân thực cao [2]. Các nghiên cứu trước đây đã chỉ ra rằng việc sử dụng GAN để tạo ảnh khuôn mặt tổng hợp có thể giúp cải thiện hiệu suất của CNN bằng cách gia tăng tính đa dạng của dữ liệu huấn luyện [17], [12]. Hơn nữa, một số biến thể của GAN như StyleGAN và CycleGAN đã được chứng minh là có thể tạo ra hình ảnh có độ chân thực cao và cải thiện độ chính xác của mô hình nhận dạng khuôn mặt [18], [5]. Ngoài ra, nghiên cứu [7] đã giới thiệu một phương pháp sử dụng GAN với cấu trúc generator nhúng residual và discriminator dựa trên Inception ResNet-V1 để cải thiện độ chính xác nhận dạng khuôn mặt trên bộ dữ liệu AR.



Hình 1. Kiến trúc tổng quát của GAN [8]

Mặc dù, các nghiên cứu gần đây đã chứng minh hiệu quả của các biến thể GAN nâng cao trong việc cải thiện hiệu suất nhận dạng khuôn mặt, việc ứng dụng GAN vẫn có giá trị thực tiễn cao trong bối cảnh nghiên cứu hiện tại. Trong bài báo này, chúng tôi đã sử dụng GAN để nâng cao hiệu suất phân loại cảm xúc khuôn mặt, từ đó tăng hiệu suất nhận dạng của mô hình CNN. Cụ thể, chúng tôi lựa chọn biến thể DCGAN để tạo sinh dữ liệu vì DCGAN tạo ra hình ảnh chất lượng khá tốt, độ phức tạp huấn luyện vừa phải, khả năng huấn luyện ổn định và phù hợp với mục tiêu tạo dữ liệu tổng hợp phục vụ cho phân loại cảm xúc khuôn mặt, một bài toán có tập dữ liệu hạn chế và phân bố mất cân bằng [13], [8]. Hơn nữa, việc sử dụng DCGAN giúp cân bằng giữa chất lượng ảnh sinh và chi phí tính toán, đặc biệt khi triển khai trên môi trường như Google Colab với tài nguyên hạn chế.

2. MẠNG ĐỐI SINH (GAN)

2.1. Hoạt động của mạng đối sinh

Mạng đối sinh là một mô hình học sâu do ông Goodfellow và các tác giả [2] đề xuất dựa trên cơ chế cạnh tranh giữa hai mạng nơ-ron:

- Mạng sinh (Generator - G): Có nhiệm vụ tạo ra dữ liệu giả sao cho trông giống dữ liệu thật nhất có thể.
- Mạng phân biệt (Discriminator - D): Có nhiệm vụ phân biệt giữa dữ liệu thật từ tập huấn luyện và dữ liệu giả do mạng sinh tạo ra.

Hai mạng này được huấn luyện đồng thời thông qua một quá trình đối kháng, nơi mạng sinh cố gắng đánh lừa mạng phân biệt, còn mạng phân biệt cố gắng phát hiện dữ liệu giả.

Trong Hình 1, Generator nhận một vector nhiễu và tạo ra một hình ảnh giả, ảnh giả đó và ảnh thật từ tập dữ liệu được đưa vào Discriminator. Sau đó, Discriminator sẽ dự đoán xem ảnh đó là thật hay giả. Nếu ảnh giả bị Discriminator phát hiện, Generator sẽ điều chỉnh lại để tạo ra ảnh giống thật hơn. Nếu Discriminator không thể phân biệt được (tức là nhận diện đó là ảnh thật), nó sẽ cải thiện bằng cách học từ dữ liệu mới. Quá trình huấn luyện sẽ được lặp lại cho đến khi Generator có thể tạo ra ảnh giả rất giống ảnh thật. Chi tiết về kiến trúc và hoạt động của Generator và Discriminator trong DCGAN được thể hiện trong Bảng 1, Bảng 2.

Bảng 1. Kiến trúc và hoạt động của Generator

Giai đoạn	Mô tả hoạt động
Đầu vào	Một vector nhiễu có kích thước nhỏ, ví dụ 100 chiều
Tầng Fully Connected	Tăng kích thước từ 100 chiều lên tensor ban đầu (ví dụ 4×4×1024)
Tầng Transposed Convolution 1	Phóng đại không gian ảnh 8×8, giảm số kênh (1024 → 512), stride=2, kernel=4.
Batch Normalization(BN)	Ổn định hóa phân phối đầu ra tầng trước
ReLU Activation	Áp dụng phi tuyến để sinh ảnh
Tầng Transposed Convolution 2	Tăng kích thước ảnh 16×16, giảm số kênh 512 → 256
BN + ReLU	Ổn định đầu ra + học tốt hơn
Tầng Transposed Convolution 3	Tăng kích thước ảnh 32×32, số kênh giảm 256 → 128
BN + ReLU	Ổn định đầu ra + học tốt hơn
Tầng Transposed Convolution cuối	Tạo ảnh RGB đầu ra (128 → 3)
Tanh Activation	Đưa giá trị pixel về [-1, 1] để phù hợp với chuẩn hóa dữ liệu
Đầu ra	Ảnh giả $x \in \mathbb{R}^{64\times64\times3x}$

Bảng 2. Kiến trúc và hoạt động của Discriminator

Giai đoạn	Mô tả hoạt động
Đầu vào	Một ảnh đầu vào $x \in \mathbb{R}^{64\times64\times3x}$, có thể là ảnh thật hoặc ảnh giả
Tầng Convolution 1	Trích đặc trưng thấp cấp (3 → 64 kênh)
Leaky ReLU (α=0.2)	Tăng khả năng học với các giá trị âm nhỏ

Tầng Convolution 2	Giảm kích thước ảnh, tăng số kênh 64 → 128
BN	Ổn định phân phối đặc trưng
LeakyReLU	Khắc phục "chết nơ-ron"
Tầng Convolution 3	Giảm kích thước ảnh, tăng số kênh 128 → 256
BN + LeakyReLU	Giúp mạng học tốt hơn, tránh chết nơ-ron
Tầng Convolution 4	Giảm kích thước ảnh, tăng số kênh 256 → 512
BN + LeakyReLU	Giúp mạng học tốt hơn, tránh chết nơ-ron
Flatten	Chuyển tensor thành vector
Tầng Fully Connected cuối	Một neuron để phân biệt thật/giả
Sigmoid	Trả xác suất ảnh là thật
Đầu ra	Giá trị D(x)∈[0,1] - dự đoán của Discriminator.

Quá trình huấn luyện GAN được thực hiện như sau:

Bước 1: Sinh ảnh giả và lấy mẫu ảnh thật. Chọn một minibatch *m* ảnh thật từ tập dữ liệu X_{real}. Sau đó tạo một minibatch ảnh giả bằng cách đưa nhiễu *z* vào Generator: $X_{fake} = G(z)$

Bước 2: Huấn luyện Discriminator. Để phân biệt ảnh thật và ảnh giả, Discriminator sẽ tối đa hoá xác suất dự đoán đúng bằng cách tính hàm mất mát (Binary Cross Entropy) [2]:

$$L_D = -\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] - \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (2.1)$$

Trong đó: - $\mathbb{E}_{x\sim p_{data}(x)}$ là kỳ vọng trên các mẫu *x* lấy từ phân phối dữ liệu thực $p_{data}(x)$.

- $D(x)$ là xác suất mà Discriminator dự đoán rằng mẫu *x* là ảnh thật, $D(x)$ gần 1 thì Discriminator tự tin rằng ảnh này là thật.
- $\mathbb{E}_{z\sim p_z(z)}$ biểu thị kỳ vọng trên các mẫu **z** lấy từ phân phối nhiễu $p_z(z)$, tức là tính trung bình trên các ảnh giả được tạo ra từ Generator.
- $G(z)$ là ảnh giả được sinh ra bởi Generator từ một vector nhiễu z.
- $D(G(z))$ là xác suất mà Discriminator dự đoán ảnh giả $G(z)$ là thật. Nếu $D(G(z)) \approx 1$ nghĩa là Discriminator bị đánh lừa bởi ảnh giả, nếu $D(G(z)) \approx 0$ là phân biệt rõ ảnh giả.
- $\log(1 - D(G(z)))$ là giá trị mất mát khi Discriminator phân biệt ảnh giả $G(z)$. Nếu $D(G(z)) \approx 1$ (tức là bị Generator đánh lừa), thì giá trị này tiến về 0, tức là Discriminator đang thất bại. nếu $D(G(z)) \approx 0$ (tức là phát hiện đúng ảnh giả), thì $\log(1 - D(G(z)))$ sẽ có giá trị lớn.

Sau đó sẽ thực hiện tối ưu hoá hàm mất mát bằng thuật toán Adam [6].

Bước 3: Huấn luyện Generator. Mục đích đánh lừa Discriminator bằng cách tối ưu hàm mất mát:

$$L_G = -\mathbb{E}_{z\sim p_z(z)}[\log D(G(z))] \quad (2.2)$$

Bước 4: Hai bước huấn luyện trên được thực hiện lặp đi lặp lại nhiều lần để cải thiện Generator và Discriminator.

2.2. Tạo bộ dữ liệu tổng hợp cho khuôn mặt người

Bộ dữ liệu hình ảnh về cảm xúc khuôn mặt được thu thập từ nền tảng Roboflow Universe (Hình 2), một hệ thống cung cấp kho dữ liệu lớn cùng các mô hình đã qua huấn luyện phục vụ cho nhận dạng hình ảnh. Nền tảng này cho phép người dùng tiếp cận hơn 575 triệu hình ảnh và hơn 750.000 bộ dữ liệu thuộc nhiều lĩnh vực khác nhau như y tế, nông nghiệp, động vật, v.v. Bên cạnh đó, người dùng có thể triển khai các mô hình đã được huấn luyện hoặc sử dụng chúng để hỗ trợ trong quá trình gán nhãn dữ liệu.

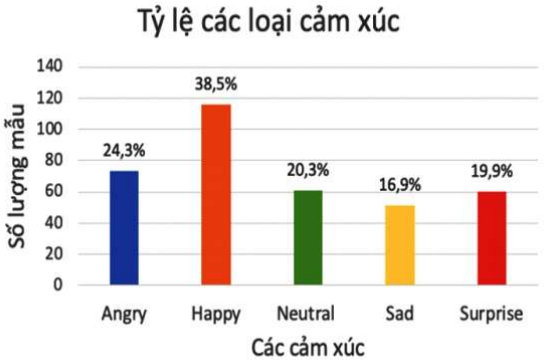
Hình 2. Một số hình ảnh trong tập dữ liệu

Bộ dữ liệu cảm xúc khuôn mặt được sử dụng trong bài báo này gồm 3.849 hình ảnh màu và được chia theo tỷ lệ 70:20:10 như sau: 2.694 ảnh cho tập huấn luyện (training), 770 ảnh cho tập đánh giá (validation) và 385 ảnh cho tập kiểm tra (test). Mỗi ảnh có kích thước 224x224 pixel (trong mô hình sẽ được chuyển về 64x64). Bộ dữ liệu được gán nhãn bao gồm các cảm xúc: Happy, Angry, Neutral, Sad và Surprise. Số lượng cụ thể và tỷ lệ được thể hiện trong Hình 3.

Chúng tôi tiến hành tạo bộ dữ liệu tổng hợp. Thay vì mỗi ảnh chỉ chứa một khuôn mặt với một cảm xúc, GAN sẽ sinh ra nhiều ảnh cho cùng một khuôn mặt và một cảm xúc, nhưng với các biến thể nhỏ (Hình 4). Từ đó, bộ dữ liệu được đưa vào nhận dạng phong phú hơn và làm tăng độ chính xác nhận dạng hơn.

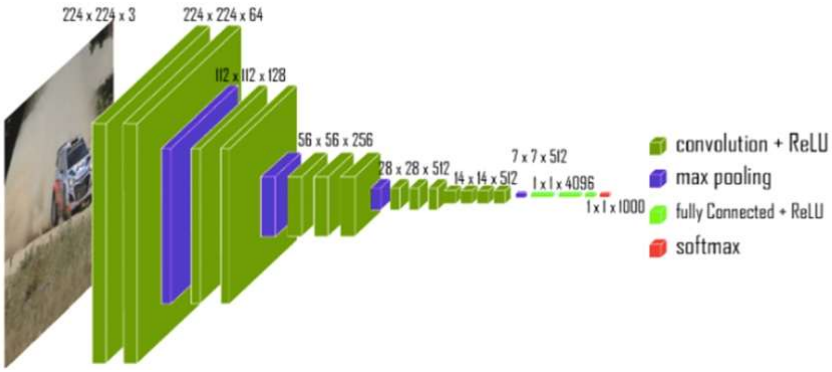
Trong bài báo này, chúng tôi sử dụng một biến thể của mô hình mạng nơ-ron tích chập CNN là mô hình VGG-16 để nhận dạng các cảm xúc khuôn mặt.

Mô hình chứa tổng cộng 16 tầng chứa trọng số, bao gồm: 13 lớp tích chập, 3 lớp kết nối đầy đủ và sử dụng hàm kích hoạt ReLU và softmax ở đầu. Mô hình thực hiện được mô tả trong Hình 5: 1) Lớp Convolution, trích xuất đặc trưng bằng các lớp tích chập nhỏ sử dụng bộ lọc 3x3, stride=1 để bộ lọc di chuyển từng bước 1, giữ nguyên độ phân giải. Sử dụng hàm kích hoạt ReLU để giảm tính phi tuyến; 2) Giảm chiều bằng lớp MaxPooling 2×2, stride = 2 giúp giảm kích thước đầu vào và giữ lại thông tin quan trọng; 3) Lớp Fully Connected (FC): Ảnh được làm phẳng thành một vector 1 chiều. Ba lớp FC gồm 2 lớp 4096 nơ-ron và 1 lớp Softmax để đưa ra kết quả dự đoán với 1000 nhãn [3].



Hình 3. Tỷ lệ các loại cảm xúc có trong dataset

Hình 4. Một số bản sao hình ảnh của các khuôn mặt sau khi áp dụng GAN



Hình 5. Mô hình VGG-16 (Nguồn: https://www.geeksforgeeks.org/vgg-16-cnn-model/)

2.3. Mô hình mạng nơ-ron tích chập

Mô hình VGG-16 gốc được huấn luyện sẵn trên tập dữ liệu ImageNet và đầu ra là 1000 nhãn ứng với các đối tượng khác nhau (ô tô, máy bay...). Tuy nhiên, khi áp dụng vào bài toán nhận dạng cảm xúc khuôn mặt chúng tôi đã tuỳ chỉnh lại mô hình trong quá trình thực nghiệm bằng cách giữ lại toàn bộ phần trích xuất đặc trưng của lớp Convolution, còn lớp FC cuối cùng sẽ được huấn luyện lại thành lớp mới có số đầu ra là 5 (5 cảm xúc).

Chúng tôi chọn mô hình VGG-16 vì nó có độ chính xác cao, dễ triển khai và phù hợp với tài nguyên tính toán của bài toán. So với các mô hình khác như AlexNet hay MobileNet, VGG-16 vượt trội về khả năng nhận diện đặc trưng [4], [1].

3. KẾT QUẢ ỨNG DỤNG GAN TRONG MÔ HÌNH VGG-16

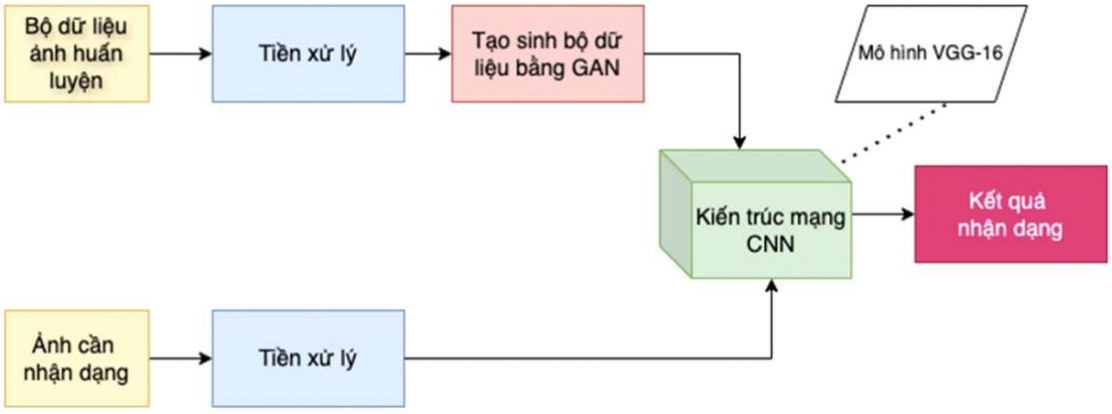
3.1. Các phương pháp đánh giá [9]

- **Confusion Matrix (ma trận nhầm lẫn)**

Ma trận nhầm lẫn là một bảng hai chiều (Bảng 3), trong đó mỗi hàng của bảng đại diện cho một lớp thực tế và mỗi cột đại diện cho một lớp dự đoán. Mỗi phần tử trong ma trận là số lượng các mẫu thuộc loại tương ứng giữa thực tế và dự đoán.

Bảng 3. Ma trận nhầm lẫn

	Dự đoán A	Dự đoán B	Dự đoán C
Thực tế = A	TP_A	FP_A_B	FP_C_C
Thực tế = B	FP_B_A	TP_B	FP_B_C
Thực tế = C	FP__A	FP_C_B	TP_C



Hình 6. Quá trình huấn luyện và dự đoán nhãn cảm xúc của ảnh

Trong đó, True Positive (TP_A, TP_B, TP_C) là số lượng mẫu được phân loại chính xác cho mỗi lớp. False Positive (FP_A_B, FP_A_C, FP_B_A,...) là số lượng mẫu được phân loại sai giữa các lớp.

- **Precision**

Precision là chỉ số đánh giá mức độ chính xác của mô hình trong việc dự đoán các mẫu thuộc lớp positive (lớp dương tính).

- **Recall**

$$R = \frac{TruePositive}{TruePositive+FalseNegative} \quad (3.2)$$

Recall là thước đo khả năng của mô hình trong việc phát hiện tất cả các mẫu thuộc lớp positive.

- **F1 score**

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (3.3)$$

F1 Score là thước đo hiệu suất dùng để đánh giá chất lượng các mô hình phân loại. Giá trị F1 Score càng cao thì mô hình hoạt động càng hiệu quả.

- **Accuracy**

$$A = \frac{Tổng\ số\ dự\ đoán\ đúng}{Tổng\ số\ mẫu} \quad (3.4)$$

Trong đó, *số lượng dự đoán đúng* là tổng số các dự đoán mà mô hình đưa ra đúng (TP và TN). *Tổng số mẫu* là tổng số tất cả các dự đoán trong bộ dữ liệu (TP, FP, FN và TN).

3.2. Kết quả đánh giá

Quá trình huấn luyện và dự đoán nhãn cảm xúc của ảnh khuôn mặt được hiển thị trong Hình 6. Trong quá trình huấn luyện, bộ ảnh huấn luyện sau bước tiền xử lý sẽ được tạo sinh thêm dữ liệu bằng GAN, sau đó sẽ được huấn luyện bằng mô hình VGG-16. Ảnh cần nhận dạng, sau bước tiền xử lý, sẽ được đưa vào mô hình để xác định cảm xúc tương ứng.

Quy trình tích hợp dữ liệu tổng hợp từ GAN vào quá trình huấn luyện mô hình VGG-16 được thực hiện cụ thể theo các bước sau:

Bước 1: Huấn luyện mô hình GAN.

Bước 2: Tạo bộ dữ liệu tổng hợp từ mô hình GAN đã huấn luyện ở bước 1. Bước này tạo ra tập các ảnh giả cho các lớp cảm xúc và được gán nhãn tương ứng.

Bước 3: Tạo tập dữ liệu huấn luyện cho mô hình VGG-16. Tập dữ liệu sẽ bao gồm dữ liệu ảnh thât và dữ liệu ảnh giả. Trong bài báo này tỷ lệ ảnh thật: ảnh giả được sử dụng là 1:1.

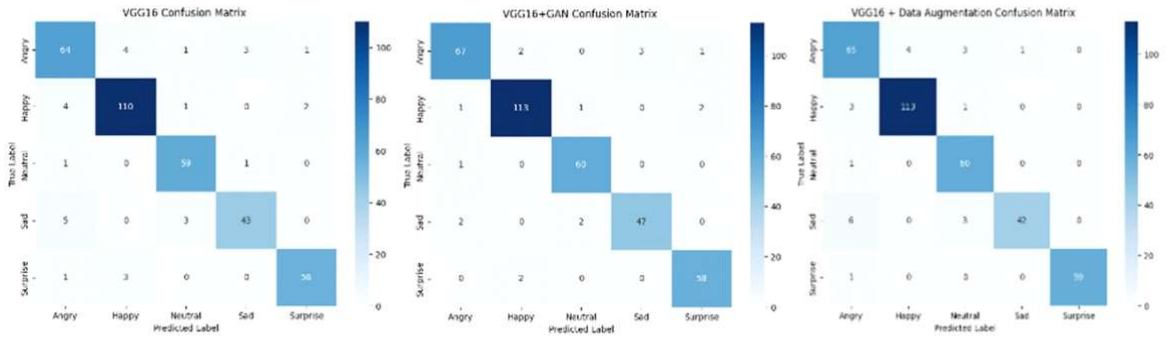
Bước 4: Huấn luyện mô hình VGG-16. Dữ liệu ảnh đầu vào của mô hình là tập dữ liệu được tạo ở bước 3.

3.2.1. Đánh giá các phương pháp đề xuất

Chúng tôi sử dụng phương pháp mạng GAN và phương pháp tăng cường dữ liệu truyền thống (TCDLTT) để tăng hiệu suất phân loại cảm xúc khuôn mặt từ đó cải thiện hiệu suất nhận dạng của mô hình.

Phương pháp tăng cường dữ liệu truyền thống sử dụng các kỹ thuật xoay ảnh, lật ảnh, điều chỉnh độ sáng và độ tương phản để mở rộng tập dữ liệu của mô hình VGG-16, từ đó có thể so sánh khách quan về tính hiệu quả của mô hình khi sử dụng phương pháp GAN với phương pháp TCDLTT.

Chúng ta có thể quan sát kết quả nhận dạng đúng, kết quả nhận dạng nhầm khi ứng dụng các phương pháp tăng cường phân loại cảm xúc trong ma trận nhầm lẫn được thể hiện ở Hình 7. Từ ma trận nhầm lẫn, chúng tôi tính được xác suất nhận dạng cảm xúc khuôn mặt của mô hình theo các phương pháp đánh giá. Hiệu suất nhận dạng của mô hình VGG-16 khi ứng dụng phương pháp TCDLTT tăng từ 91,7% lên 93,6%, trong khi đó hiệu suất nhận dạng của mô hình khi ứng dụng phương pháp GAN tăng từ 91,7% lên 95,3% (theo phương pháp đánh giá Accuracy). Kết quả nhận dạng của các phương pháp được thể hiện trong Hình 8.



a) b) c)

Hình 7. Ma trận nhầm lẫn của mô hình VGG-16 sử dụng các phương pháp để tăng phân loại cảm xúc khuôn mặt: a) Không sử dụng; b) Phương pháp GAN; c) Phương pháp tăng cường dữ liệu

Bảng 4. Bảng so sánh kết quả khi sử dụng phương pháp GAN và TCDLTT.

Tiêu chí so sánh	GAN	TCDLTT
Thời gian huấn luyện	2h15’	1h25’
Chi phí tính toán	0	0
Kết quả nhận dạng	95,3%	93,6%
Yêu cầu phần cứng	Cần có CPU mạnh	Có CPU là đủ

Bảng 4 thể hiện kết quả so sánh giữa phương pháp sử dụng mạng GAN và phương pháp TCDLTT. Trong bài báo này, chúng tôi thực hiện lập trình trên Google Colab để bảo đảm chương trình chạy ổn định vì vậy Chi phí tính toán là 0. Với epoch là 100, Batch size 64, ảnh kích thước 64x64, thì thời gian huấn luyện của GAN sẽ dài hơn so với phương pháp tăng cường dữ liệu truyền thống và phương pháp GAN đòi hỏi yêu cầu phần cứng cao hơn TCDLTT, tuy nhiên kết quả nhận dạng khi sử dụng GAN cao hơn. Chính vì vậy, sử dụng phương pháp GAN để tăng cường dữ liệu sẽ nâng cao được hiệu suất nhận dạng cao hơn

```
HIỆU SUẤT NHẬN DẠNG CỦA VGG16
Accuracy: 0.9171 (91.71%)
Angry :
  Precision: 0.8533
  Recall: 0.8767
  F1-Score: 0.8649

Happy :
  Precision: 0.9402
  Recall: 0.9402
  F1-Score: 0.9402

Neutral :
  Precision: 0.9219
  Recall: 0.9672
  F1-Score: 0.9440

Sad :
  Precision: 0.9149
  Recall: 0.8431
  F1-Score: 0.8776

Surprise :
  Precision: 0.9492
  Recall: 0.9333
  F1-Score: 0.9412
```

```
HIỆU SUẤT NHẬN DẠNG CỦA VGG16+GAN
Accuracy: 0.9530 (95.30%)
Angry :
  Precision: 0.9437
  Recall: 0.9178
  F1-Score: 0.9306

Happy :
  Precision: 0.9658
  Recall: 0.9658
  F1-Score: 0.9658

Neutral :
  Precision: 0.9524
  Recall: 0.9836
  F1-Score: 0.9677

Sad :
  Precision: 0.9400
  Recall: 0.9216
  F1-Score: 0.9307

Surprise :
  Precision: 0.9508
  Recall: 0.9667
  F1-Score: 0.9587
```

```
HIỆU SUẤT NHẬN DẠNG CỦA VGG16+Data Augmentation
Accuracy: 0.9365 (93.65%)
Angry :
  Precision: 0.8553
  Recall: 0.8904
  F1-Score: 0.8725

Happy :
  Precision: 0.9658
  Recall: 0.9658
  F1-Score: 0.9658

Neutral :
  Precision: 0.8955
  Recall: 0.9836
  F1-Score: 0.9375

Sad :
  Precision: 0.9767
  Recall: 0.8235
  F1-Score: 0.8936

Surprise :
  Precision: 1.0000
  Recall: 0.9833
  F1-Score: 0.9916
```

Hình 8. Hiệu suất nhận dạng của mô hình VGG-16 áp dụng các phương pháp tăng cường phân loại cảm xúc

phương pháp TCDLTT.

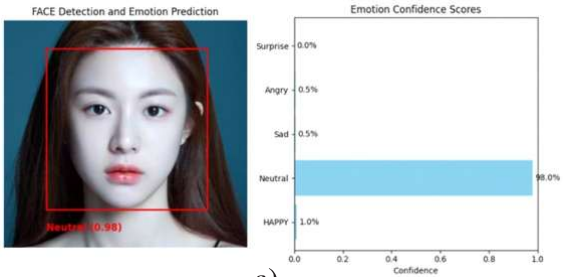
3.2.2. Dự đoán trên dữ liệu thực tế

Kết quả nhận dạng cảm xúc với các ảnh thực tế được hiển thị trong Hình 9. Đối với ảnh có nhiều khuôn mặt, mô hình cũng thực hiện việc nhận dạng từng khuôn mặt và cho ra kết quả nhận dạng.

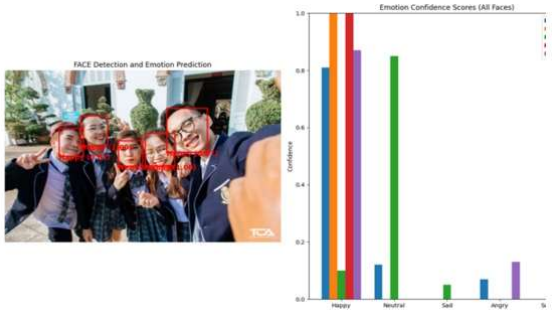
Các bước tiền xử lý ảnh đầu vào của dữ liệu thực tế bao gồm:

- Chuyển đổi ảnh đầu vào từ RGB sang mảng NumPy.
- Chuyển ảnh màu thành ảnh xám để phát hiện khuôn mặt.
- Phát hiện và cắt khu vực chứa khuôn mặt.
- Resize khuôn mặt thành kích thước cố định phù hợp với input của mô hình.
- Chuyển đổi sang mảng NumPy, thêm chiều batch, và chuẩn hóa giá trị pixel.

Sau bước tiền xử lý là bước trích xuất đặc trưng của khuôn mặt, feature map sẽ được tạo ra ở các lớp convolution của mô hình (Hình 10). Và các feature map tiếp tục được xử lý ở các tầng sâu hơn để cuối cùng mô hình sẽ dự đoán nhãn cảm xúc.



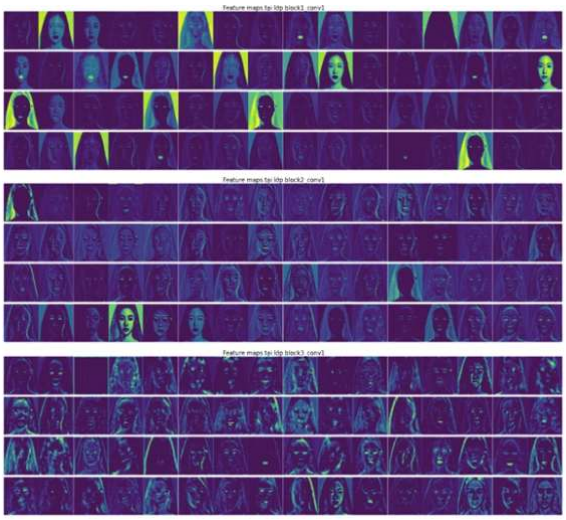
a)



b)

Hình 9. Kết quả dự đoán trên dữ liệu thực tế: a) Ảnh chân dung, b) Ảnh nhiều khuôn mặt

Các tham số và giá trị các tham số được sử dụng trong chương trình được thể hiện trong Bảng 5.



Hình 10. Feature Map tại các lớp

Bảng 5. Các tham số được sử dụng trong chương trình

Tham số	Giá trị	Mô tả
latent_dim	Kích thước vector nhiễu đầu vào cho Generator	100
num_classes	Số lớp cảm xúc phân loại	5
device	Thiết bị sử dụng	"cuda"
batch_size	Số lượng ảnh xử lý mỗi batch	64
epochs_gan	Số vòng lặp huấn luyện cho GAN	100
epochs_vgg	Số vòng lặp huấn luyện cho mô hình VGG16	30
lr_gan	Tốc độ học cho GAN	0.0002
lr_vgg	Tốc độ học cho VGG16	0.001
beta1, beta2	Tham số beta cho optimizer Adam trong GAN	0.5, 0.999
transform	Chuỗi biến đổi ảnh (resize, normalize)	64x64, ([0.5])
loss_fn_gan	Hàm mất mát trong GAN	BCELoss
loss_fn_vgg	Hàm mất mát trong VGG16	CrossEntropy Loss

4. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất một phương pháp để nâng cao hiệu suất nhận dạng của mô hình VGG-16 bằng cách sử dụng mạng đối sinh GAN để tạo bộ dữ liệu tổng hợp cho mô hình. Phương pháp này giúp cho tập huấn luyện trở nên đa dạng hơn và làm tăng hiệu suất nhận dạng từ 91,7% lên 95,3% theo phương pháp đánh giá Accuracy. Độ chính xác đạt đến 95,3% chứng tỏ phương pháp này giúp cho mô hình học được nhiều đặc trưng tốt hơn từ dữ liệu. Mô hình đạt độ chính xác cao, tuy nhiên vẫn còn một số nhận dạng nhầm lẫn giữa các mẫu, đặc biệt là giữa "Angry" và "Sad". Sự nhầm lẫn này là do sự tương đồng về khuôn mặt, thường cả hai cảm xúc "Angry" và "Sad" có lông mày hướng xuống và nhíu lại, ánh mắt buồn và tập trung, không có nhiều biểu hiện tích cực. Điều này sẽ dẫn đến sự chồng lặp về đặc trưng hình ảnh và sẽ dẫn đến sự nhầm lẫn.

Phương pháp ứng dụng mạng GAN để nâng cao hiệu suất phân loại cảm xúc khuôn mặt có thể được ứng dụng trong y tế để phân loại ảnh X-Quang, MRI vì dữ liệu có nhãn rất hiếm, hoặc áp dụng cho nhận dạng khuôn mặt từ camera an ninh trong điều kiện ánh sáng kém. Vì GAN có thể sinh ảnh giả mang đặc trưng phong phú sẽ giúp mô hình học được nhiều đặc điểm hơn từ dữ liệu gốc.

TÀI LIỆU THAM KHẢO

[1] Du, X., Sun, Y., Song, Y., Sun, H., & Yang, L. (2023), "A comparative study of different CNN models and transfer learning effect for underwater object classification in side-scan sonar images", *Remote Sensing*, 15(3), 593.

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y.. (2014), "Generative Adversarial Networks", *arXiv preprint arXiv*:1406.2661.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016), *"Deep residual learning for image recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*,pp. 770-778.

[4] Hindarto, D. (2023), "Comparative analysis vgg16 vs mobilenet performance for fish identification", *International Journal Software Engineering and Computer Science (IJSECS)*, 3(3), 270-280.

[5] Jabberi, M., Wali, A., & Alimi, A. M. (2023), "Generative data augmentation applied to face recognition", *In 2023 International Conference on Information Networking (ICOIN)*, IEEE, *pp. 242-247.*

[6] Kingma, D. P., & Ba, J. (2014), "Adam: A Method for Stochastic Optimization", *arXiv*:1412.6980.

[7] Li, Z., Li, Z., & Li, X. (2025), "Facial Recognition Leveraging Generative Adversarial Networks", *arXiv preprint arXiv:2505.11884.*

[8] Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., & Malossi, C. (2018), "Bagan: Data augmentation with balancing GAN", *arXiv preprint arXiv*:1803.09655.

[9] Müller, Andreas C., and Sarah Guido (2016), *Introduction to machine learning with Python: a guide for data scientists*, O'Reilly Media.

[10] Nemavhola, A., Chibaya, C., & Viriri, S. (2025), "A Systematic Review of CNN Architectures, Databases, Performance Metrics, and Applications in Face Recognition", *Information*, 16(2), 107.

[11] Perez, L., & Wang, J. (2017), "The effectiveness of data augmentation in image classification using deep learning", *arXiv preprint arXiv*:1712.04621.

[12] Qiu, R., Zhao, P., & Zhu, J. (2021), "Face Data Augmentation using Generative Adversarial Networks for Improved Face Recognition Performance", *IEEE Access*, 9, pp. 13424-13436.

[13] Radford, A., Metz, L., & Chintala, S. (2015), "Unsupervised representation learning with deep convolutional generative adversarial networks", *arXiv preprint arXiv*:1511.06434.

[14] Shorten, C., & Khoshgoftaar, T. M. (2019), "A survey on image data augmentation for deep learning", *Journal of Big Data*, 6(1), pp. 1-48.

[15] Wang, M., & Deng, W. (2021), "Deep face recognition: A survey", *Neurocomputing*, 429, 215-244.

[16] Wang, X., Peng, J., Zhang, S., Chen, B., Wang, Y., & Guo, Y. (2022), "A survey of face recognition", *arXiv preprint arXiv*: 2212.13038.

[17] Yi, H., Zhang, J., & Zhang, Z. (2019), "DualGAN-based Data Augmentation for Face Recognition", *Pattern Recognition Letters*, 128, pp. 242-248.

[18] Zhao, Y., Lin, H., Zhang, Y., & Wang, C. (2021), "Enhancing Face Recognition Performance with GAN-based Augmented Data", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2), pp. 182-195.

[19] Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017), "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", *In Proceedings of the IEEE international conference on computer vision*, pp. 2223-2232.

Liên hệ:

TS. Nguyễn Thị Hà Phương

Khoa Kỹ thuật và Công nghệ, Đại học Huế

Địa chỉ: Số 1 Điện Biên Phủ, Quận Thuận Hoá, Thành phố Huế

Email: nthphuong.huet@hueuni.edu.vn

Ngày nhận bài: 13/5/2025

Ngày gửi phản biện: 13/5/2025

Ngày duyệt đăng: 15/8/2025