ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

**KHOA CÔNG NGHỆ THÔNG TIN**



**PHÂN TÍCH LUẬN VĂN NHẬN DIỆN CẢM XÚC MẶT NGƯỜI**

**SỬ DỤNG MẠNG HỌC SÂU CÓ CHÚ Ý**

# 

**Giảng viên hướng dẫn:** Đỗ Như Tài

**Sinh viên thực hiện:** Đặng Huy Hoàng – 3122560019

Đặng Huy Hoàng – 3122560020

**TP. HỒ CHÍ MINH, tháng 4 năm 2025**

[**1. Tên đồ án**](#_owmlpdk4fd7y) **3**

[1.1. Ý nghĩa của tiêu đề](#_twth1qhtdt4m) 3

[1.2. Bối cảnh và tầm quan trọng của bài toán](#_28a3hz1a6bn0) 4

[1.3. Định hướng kỹ thuật của luận văn](#_ubxixudhen84) 4

[1.4. Kết nối với các nghiên cứu liên quan](#_e7tcx9vfq3gz) 5

[**2. Tác giả, Nơi công bố, Năm công bố**](#_7b8qxydbdxvv) **5**

[2.1. Tác giả](#_bjz08zd2medf) 5

[2.2. Nơi công bố](#_hcazhkdhx5q9) 5

[2.3. Năm công bố](#_b95e0t29fwc0) 6

[2.4. Sự hỗ trợ khoa học](#_xqzkqqeqoa8p) 6

[**3. Mục tiêu nghiên cứu**](#_785pcqiwxe3q) **7**

[3.1. Ứng dụng Residual Masking Network để phân loại biểu cảm cảm xúc](#_p5n80jyfc8oq) 7

[3.2. Xây dựng và phát triển tập dữ liệu VEMO](#_t9h6njsytx1s) 7

[3.3. Tăng cường độ chính xác bằng phương pháp học kết hợp](#_lcyx2jx7bbv0) 8

[3.4. Đánh giá hiệu quả và so sánh với các nghiên cứu trước](#_45ukgosbtbo7) 8

[**4. Cơ sở lý thuyết**](#_212ega669n4x) **10**

[4.1. Nhận diện cảm xúc khuôn mặt](#_nbgnlu5s3qqt) 10

[4.2. Mạng nơ-ron tích chập (CNN)](#_hrd5yo8vwtm) 11

[4.3. Cơ chế chú ý (Attention Mechanism)](#_nutq9njh4jp6) 11

[4.4. Tập dữ liệu chuẩn](#_h88ylmj5izi1) 12

[4.5. Các phương pháp trực quan hóa](#_z7squns4uow6) 13

[**5. Phương pháp thu thập và xử lý thông tin**](#_i3zcbdvp4tsn) **14**

[5.1. Thu thập dữ liệu](#_ouhfzm5u2jr6) 14

[5.2. Xử lý dữ liệu](#_mrpfbmjjak2m) 15

[5.3. Phương pháp thực nghiệm](#_2o0bpy327o3r) 16

[5.4. Trực quan hóa kết quả](#_qzyxhgxm8igj) 18

[**6. Kết quả đạt được**](#_58rqwtsbc50y) **19**

[6.1. Hiệu quả của Residual Masking Network](#_fb11bxdnbu8i) 19

[6.2. Hiệu quả trên tập dữ liệu VEMO](#_pbmogf3prszi) 20

[6.3. Đóng góp của tập dữ liệu VEMO](#_u7l7l7fjln80) 21

[6.4. Trực quan hóa và giải thích mô hình](#_9b013vpq2mee) 21

[6.5. Ứng dụng thực tế](#_uboyfij8rbxq) 22

[**7. Hạn chế của công trình**](#_85nnjctmhk66) **23**

[7.1. Thời gian nghiên cứu hạn chế](#_knp0hoxxbvht) 24

[7.2. Số lượng tham số lớn](#_ozotgk1zw8o6) 24

[7.3. Chưa kiểm chứng trên các bài toán khác](#_9r5ez6jmxj05) 25

[7.4. Hạn chế của tập dữ liệu VEMO](#_vv7mhqox7eyk) 26

[7.5. Khả năng giải thích mô hình](#_2tep06pms0dn) 26

[**8. Kết luận**](#_3qhfzqyamcda) **28**

# **Tên đồ án**

Luận văn mang tiêu đề **“Nhận diện cảm xúc mặt người sử dụng mạng học sâu có cơ chế chú ý”** là một công trình nghiên cứu chuyên sâu trong lĩnh vực trí tuệ nhân tạo (AI) và thị giác máy tính (computer vision). Tên gọi của đồ án không chỉ phản ánh trọng tâm khoa học mà còn gợi mở những khía cạnh kỹ thuật tiên tiến được tác giả khai thác để giải quyết bài toán nhận diện cảm xúc con người – một trong những thách thức quan trọng của AI hiện đại. Công trình này tập trung vào việc ứng dụng các kỹ thuật **học sâu (deep learning)**, cụ thể là **mạng nơ-ron tích chập (Convolutional Neural Network - CNN)**, kết hợp với **cơ chế chú ý (attention mechanism)**, nhằm phân loại chính xác các biểu cảm cảm xúc trên khuôn mặt con người, từ đó mở ra tiềm năng ứng dụng trong nhiều lĩnh vực thực tiễn như giao tiếp người-máy, giám sát tâm lý, giáo dục, và an ninh.

## **Ý nghĩa của tiêu đề**

Tiêu đề luận văn được xây dựng một cách cô đọng nhưng chứa đựng đầy đủ các yếu tố cốt lõi của nghiên cứu. Cụm từ **“Nhận diện cảm xúc mặt người”** nhấn mạnh mục tiêu chính là phân tích và hiểu các trạng thái cảm xúc được thể hiện qua biểu cảm khuôn mặt – một kênh giao tiếp phi ngôn ngữ quan trọng của con người. Đây là một bài toán phức tạp bởi cảm xúc không chỉ phụ thuộc vào các đặc điểm hình học của khuôn mặt (như vị trí mắt, miệng, lông mày) mà còn chịu ảnh hưởng từ ngữ cảnh văn hóa, sắc thái cá nhân, và các yếu tố môi trường như ánh sáng hay góc chụp.

Cụm từ **“sử dụng mạng học sâu”** chỉ ra phương pháp cốt lõi được áp dụng trong nghiên cứu. Học sâu, với nền tảng là các mạng nơ-ron nhân tạo, đã trở thành công cụ mạnh mẽ trong việc xử lý dữ liệu hình ảnh. Trong luận văn, mạng nơ-ron tích chập được sử dụng để tự động trích xuất các đặc trưng (features) từ hình ảnh khuôn mặt, thay vì dựa vào các phương pháp thủ công truyền thống như trích xuất đặc trưng dựa trên Hệ thống mã hóa cơ mặt (FACS). Điều này cho phép mô hình xử lý dữ liệu phức tạp và đa dạng hơn, đồng thời đạt được hiệu quả cao trong việc nhận diện các mẫu hình cảm xúc.

Phần cuối của tiêu đề, **“có cơ chế chú ý”**, là điểm nhấn quan trọng, phản ánh sự đột phá trong cách tiếp cận của luận văn. Cơ chế chú ý, lấy cảm hứng từ cách con người tập trung vào các chi tiết quan trọng khi quan sát, được tích hợp vào mạng học sâu để tăng cường khả năng phân tích. Thay vì xử lý toàn bộ hình ảnh khuôn mặt một cách đồng đều, cơ chế này giúp mô hình ưu tiên các vùng quan trọng (như mắt, miệng, hoặc mũi) – những khu vực thường mang thông tin cảm xúc rõ rệt nhất. Trong luận văn, cơ chế chú ý được thể hiện qua kiến trúc **Residual Masking Network**, một mô hình được thiết kế đặc biệt để tối ưu hóa việc phân loại cảm xúc.

## **Bối cảnh và tầm quan trọng của bài toán**

Bài toán nhận diện cảm xúc khuôn mặt có ý nghĩa khoa học và thực tiễn sâu rộng. Về mặt khoa học, nó đóng góp vào việc hiểu rõ hơn cách máy tính có thể mô phỏng và diễn giải các tín hiệu cảm xúc của con người, từ đó thúc đẩy sự phát triển của các hệ thống AI có khả năng giao tiếp tự nhiên hơn. Về mặt thực tiễn, nhận diện cảm xúc có thể được ứng dụng trong nhiều lĩnh vực:

* **Giao tiếp người-máy**: Giúp các trợ lý ảo hoặc robot hiểu được trạng thái cảm xúc của người dùng, từ đó đưa ra phản hồi phù hợp.
* **Y tế và tâm lý học**: Hỗ trợ chẩn đoán các rối loạn cảm xúc hoặc theo dõi trạng thái tâm lý của bệnh nhân.
* **Giáo dục**: Cá nhân hóa trải nghiệm học tập bằng cách nhận biết cảm xúc của học sinh trong quá trình học.
* **An ninh và giám sát**: Phát hiện các hành vi bất thường dựa trên biểu cảm khuôn mặt trong các hệ thống camera giám sát.

Tuy nhiên, bài toán này cũng đối mặt với nhiều thách thức, như sự khác biệt về văn hóa trong cách thể hiện cảm xúc, độ phức tạp của các biểu cảm hỗn hợp (mixed emotions), và sự ảnh hưởng của các yếu tố ngoại cảnh như ánh sáng, góc chụp, hoặc chất lượng hình ảnh. Luận văn của Phạm Quý Luận đã chọn cách tiếp cận tiên tiến bằng cách kết hợp học sâu và cơ chế chú ý để vượt qua những thách thức này, đồng thời bổ sung một tập dữ liệu mới (VEMO) dành riêng cho người Việt Nam, góp phần làm phong phú thêm tài nguyên nghiên cứu trong nước.

## **Định hướng kỹ thuật của luận văn**

Luận văn tập trung vào việc phát triển và đánh giá một kiến trúc mạng học sâu mới, được gọi là **Residual Masking Network**, được thiết kế để tích hợp cơ chế chú ý vào quá trình trích xuất đặc trưng. Không giống các mô hình CNN truyền thống xử lý toàn bộ hình ảnh một cách đồng nhất, Residual Masking Network sử dụng các kỹ thuật “masking” (che phủ) để tập trung vào các vùng quan trọng của khuôn mặt, từ đó nâng cao độ chính xác trong phân loại cảm xúc. Mô hình này được huấn luyện trên cả tập dữ liệu công khai **FER2013** (chứa hơn 35,000 hình ảnh khuôn mặt với bảy loại cảm xúc) và tập dữ liệu **VEMO** do chính tác giả xây dựng, tập trung vào đặc điểm khuôn mặt của người Việt Nam.

Ngoài ra, luận văn còn khám phá phương pháp **học kết hợp (ensemble learning)**, kết hợp Residual Masking Network với các mô hình CNN khác để cải thiện hiệu suất. Việc sử dụng cơ chế chú ý và học kết hợp thể hiện nỗ lực của tác giả trong việc không chỉ nâng cao độ chính xác mà còn tăng tính tổng quát hóa của mô hình trên các tập dữ liệu khác nhau.

## **Kết nối với các nghiên cứu liên quan**

Tiêu đề luận văn cũng phản ánh mối liên hệ chặt chẽ với các nghiên cứu trước đó trong lĩnh vực nhận diện cảm xúc. Tác giả đã tham khảo các công trình của Paul Ekman về biểu cảm cảm xúc cơ bản, cũng như các kiến trúc mạng học sâu hiện đại như ResNet, VGG, và Inception. Tuy nhiên, bằng cách tích hợp cơ chế chú ý và xây dựng tập dữ liệu VEMO, luận văn mang đến một góc nhìn mới, đặc biệt phù hợp với bối cảnh Việt Nam, nơi các nghiên cứu về nhận diện cảm xúc còn hạn chế về dữ liệu và tài nguyên.

# **Tác giả, Nơi công bố, Năm công bố**

## **Tác giả**

Luận văn được thực hiện bởi **Phạm Quý Luận**, một sinh viên chuyên ngành **Khoa học Máy tính** tại **Trường Đại học Bách Khoa**, trực thuộc **Đại học Quốc gia Thành phố Hồ Chí Minh**. Là một sinh viên đại học, Phạm Quý Luận đã thể hiện năng lực nghiên cứu đáng chú ý thông qua việc lựa chọn một bài toán phức tạp trong lĩnh vực trí tuệ nhân tạo (AI) – nhận diện cảm xúc khuôn mặt – và ứng dụng các kỹ thuật học sâu tiên tiến để giải quyết. Sự lựa chọn này không chỉ phản ánh sự đam mê với khoa học máy tính mà còn cho thấy khả năng tiếp cận các phương pháp nghiên cứu hiện đại của tác giả. Trong luận văn, Phạm Quý Luận đã tự cam kết rằng đây là công trình nghiên cứu độc lập dưới sự hướng dẫn khoa học, đồng thời chịu hoàn toàn trách nhiệm về tính trung thực và bản quyền của nội dung.

## **Nơi công bố**

Luận văn được thực hiện và bảo vệ tại **Khoa Khoa học - Kỹ thuật Máy tính**, thuộc **Trường Đại học Bách Khoa**, một trong những đơn vị đào tạo và nghiên cứu hàng đầu về công nghệ và kỹ thuật tại Việt Nam, trực thuộc **Đại học Quốc gia Thành phố Hồ Chí Minh**. Trường Đại học Bách Khoa nổi tiếng với môi trường học thuật nghiêm túc, đội ngũ giảng viên giàu kinh nghiệm, và cơ sở vật chất hiện đại, tạo điều kiện thuận lợi cho các nghiên cứu ứng dụng trong lĩnh vực công nghệ thông tin, đặc biệt là trí tuệ nhân tạo và học máy. Khoa Khoa học - Kỹ thuật Máy tính, nơi luận văn được triển khai, là một trung tâm nghiên cứu và đào tạo chuyên sâu về các lĩnh vực như học sâu, xử lý hình ảnh, và thị giác máy tính – những nền tảng cốt lõi của công trình này. Việc bảo vệ luận văn tại đây đảm bảo rằng nghiên cứu đã trải qua quá trình đánh giá nghiêm ngặt bởi hội đồng chuyên môn, khẳng định tính học thuật và giá trị khoa học của công trình.

## **Năm công bố**

Luận văn được hoàn thành và công bố vào **tháng 12 năm 2019**. Đây là thời điểm mà lĩnh vực học sâu đang phát triển mạnh mẽ trên toàn cầu, với nhiều tiến bộ trong các kiến trúc mạng nơ-ron như ResNet, Inception, và các cơ chế chú ý (attention mechanisms). Tại Việt Nam, các nghiên cứu về nhận diện cảm xúc khuôn mặt vẫn còn ở giai đoạn sơ khai, đặc biệt là về dữ liệu và ứng dụng thực tiễn. Việc công bố luận văn vào thời điểm này không chỉ thể hiện sự nhạy bén của tác giả trong việc nắm bắt xu hướng công nghệ mà còn đánh dấu một đóng góp quan trọng vào việc phát triển các nghiên cứu AI tại Việt Nam, đặc biệt thông qua việc xây dựng tập dữ liệu VEMO dành riêng cho người Việt Nam.

## **Sự hỗ trợ khoa học**

Luận văn được thực hiện dưới sự hướng dẫn của **Tiến sĩ Trần Tuấn Anh**, một giảng viên giàu kinh nghiệm tại Trường Đại học Bách Khoa. Vai trò của Tiến sĩ Trần Tuấn Anh không chỉ dừng lại ở việc định hướng nghiên cứu mà còn bao gồm hỗ trợ tác giả trong việc tiếp cận các phương pháp học sâu hiện đại, thiết kế thí nghiệm, và đánh giá kết quả. Sự hướng dẫn này đã góp phần đảm bảo tính chặt chẽ và khoa học của công trình, từ việc lựa chọn kiến trúc mạng Residual Masking Network đến việc xây dựng tập dữ liệu mới.

Bên cạnh đó, luận văn được phản biện bởi **Tiến sĩ Nguyễn Hồ Mẫn Răng**, một chuyên gia trong lĩnh vực khoa học máy tính. Quá trình phản biện đảm bảo rằng các phương pháp, kết quả, và kết luận của luận văn được đánh giá một cách khách quan và toàn diện, từ đó nâng cao độ tin cậy và chất lượng của nghiên cứu. Sự tham gia của hai tiến sĩ này không chỉ khẳng định tính nghiêm túc của luận văn mà còn phản ánh sự đầu tư về mặt học thuật từ phía Khoa Khoa học - Kỹ thuật Máy tính.

# **Mục tiêu nghiên cứu**

Mục tiêu chính của luận văn **“Nhận diện cảm xúc mặt người sử dụng mạng học sâu có cơ chế chú ý”** là phát triển một hệ thống nhận diện cảm xúc trên khuôn mặt con người với độ chính xác cao, tận dụng các kỹ thuật học sâu (deep learning) tiên tiến để giải quyết bài toán phức tạp này. Nghiên cứu không chỉ hướng đến việc cải thiện hiệu suất kỹ thuật mà còn mang tầm nhìn rộng hơn, góp phần vào sự phát triển của trí tuệ nhân tạo (AI) tại Việt Nam thông qua việc xây dựng tài nguyên dữ liệu mới và mở ra các ứng dụng thực tiễn trong nhiều lĩnh vực. Cụ thể, luận văn đặt ra bốn mục tiêu trọng tâm, được kết hợp chặt chẽ để tạo nên một công trình nghiên cứu toàn diện, cùng với sự nhấn mạnh vào ý nghĩa thực tiễn của các kết quả đạt được.

## **Ứng dụng Residual Masking Network để phân loại biểu cảm cảm xúc**

Mục tiêu đầu tiên của luận văn là thiết kế và ứng dụng **Residual Masking Network**, một kiến trúc mạng học sâu được tích hợp **cơ chế chú ý (attention mechanism)**, nhằm phân loại các biểu cảm cảm xúc dựa trên hình ảnh khuôn mặt. Residual Masking Network được phát triển để khắc phục hạn chế của các mô hình truyền thống, vốn thường xử lý toàn bộ hình ảnh một cách đồng đều mà không tập trung vào các vùng quan trọng mang thông tin cảm xúc (như mắt, miệng, hoặc lông mày).

Cơ chế chú ý trong Residual Masking Network cho phép mô hình ưu tiên phân tích các vùng đặc trưng trên khuôn mặt, từ đó nâng cao độ chính xác trong việc nhận diện các cảm xúc như giận dữ, hạnh phúc, buồn bã, sợ hãi, ghê tởm, ngạc nhiên, và trạng thái trung lập. Việc sử dụng mô hình này không chỉ thể hiện sự đổi mới kỹ thuật mà còn phản ánh nỗ lực của tác giả trong việc tích hợp các tiến bộ mới nhất của học sâu, đặc biệt là các cơ chế chú ý vốn đang trở thành xu hướng chủ đạo trong xử lý hình ảnh và thị giác máy tính. Mục tiêu này đặt nền tảng cho các thí nghiệm thực nghiệm, giúp đánh giá hiệu quả của một kiến trúc mạng tùy chỉnh so với các mô hình tiêu chuẩn như ResNet, VGG, hay Inception.

## **Xây dựng và phát triển tập dữ liệu VEMO**

Một mục tiêu quan trọng khác của luận văn là xây dựng và phát triển **tập dữ liệu VEMO**, chứa các hình ảnh khuôn mặt của người Việt Nam, nhằm đóng góp trực tiếp vào nghiên cứu bài toán nhận diện cảm xúc tại Việt Nam. Trong bối cảnh các tập dữ liệu quốc tế như FER2013 hay AffectNet chủ yếu bao gồm hình ảnh từ các quốc gia phương Tây, sự thiếu hụt dữ liệu đại diện cho đặc điểm khuôn mặt và cách thể hiện cảm xúc của người Việt Nam là một rào cản lớn đối với các nghiên cứu địa phương.

Tập dữ liệu VEMO được thiết kế để phản ánh các đặc điểm văn hóa, sắc thái cảm xúc, và sự đa dạng về biểu cảm của người Việt Nam, từ đó tạo ra một nguồn tài nguyên quý giá cho các nghiên cứu tiếp theo. Mục tiêu này không chỉ mang ý nghĩa kỹ thuật mà còn thể hiện tầm nhìn chiến lược của tác giả trong việc thúc đẩy sự phát triển của AI tại Việt Nam, đặc biệt trong các bài toán liên quan đến xử lý tín hiệu con người. Việc xây dựng VEMO cũng đặt ra thách thức về thu thập, xử lý, và gắn nhãn dữ liệu, nhưng đồng thời khẳng định cam kết của luận văn trong việc tạo ra giá trị bền vững cho cộng đồng nghiên cứu.

## **Tăng cường độ chính xác bằng phương pháp học kết hợp**

Mục tiêu thứ ba của luận văn là nâng cao độ chính xác của hệ thống nhận diện cảm xúc thông qua việc sử dụng **phương pháp học kết hợp (ensemble learning)**, kết hợp **Residual Masking Network** với nhiều mô hình học sâu hiện đại khác. Học kết hợp là một kỹ thuật tiên tiến, cho phép tận dụng sức mạnh của nhiều mô hình để bù đắp điểm yếu của từng mô hình riêng lẻ, từ đó cải thiện hiệu suất tổng thể.

Trong luận văn, tác giả đã tích hợp Residual Masking Network với sáu mô hình CNN khác, chẳng hạn như ResNet, DenseNet, hoặc Inception, để tạo ra một hệ thống mạnh mẽ hơn. Mục tiêu này không chỉ nhằm tối ưu hóa độ chính xác mà còn hướng đến việc đảm bảo tính tổng quát hóa (generalization) của mô hình trên các tập dữ liệu khác nhau, bao gồm cả dữ liệu quốc tế (FER2013) và dữ liệu địa phương (VEMO). Việc áp dụng học kết hợp thể hiện sự hiểu biết sâu sắc của tác giả về các kỹ thuật học máy tiên tiến và khả năng sáng tạo trong việc kết hợp chúng để giải quyết bài toán thực tiễn.

## **Đánh giá hiệu quả và so sánh với các nghiên cứu trước**

Mục tiêu thứ tư là **đánh giá hiệu quả của các mô hình đề xuất** trên hai tập dữ liệu chính – **FER2013** và **VEMO** – đồng thời so sánh kết quả với các nghiên cứu khoa học trước đó để khẳng định giá trị của công trình. Tập dữ liệu FER2013, với hơn 35,000 hình ảnh khuôn mặt và bảy loại cảm xúc, là một chuẩn mực quốc tế, cho phép đánh giá khả năng cạnh tranh của mô hình trên phạm vi toàn cầu. Trong khi đó, tập dữ liệu VEMO cung cấp cơ hội để kiểm tra hiệu quả của mô hình trong bối cảnh đặc thù của người Việt Nam.

Quá trình đánh giá được thực hiện thông qua các chỉ số như **độ chính xác (accuracy)** và **ma trận bối rối (confusion matrix)**, giúp phân tích chi tiết khả năng phân loại từng loại cảm xúc. Tác giả cũng so sánh hiệu suất của Residual Masking Network và phương pháp học kết hợp với các kiến trúc hiện đại (như VGG19, ResNet18, DenseNet121) và các nghiên cứu được công bố trước đó (như Deep-Emotion, Ensemble MLCNNs). Mục tiêu này không chỉ nhằm xác nhận tính ưu việt của các phương pháp đề xuất mà còn cung cấp một cơ sở tham khảo đáng tin cậy cho các nghiên cứu tiếp theo trong lĩnh vực nhận diện cảm xúc.

Ngoài các mục tiêu kỹ thuật, luận văn nhấn mạnh mạnh mẽ vào **ý nghĩa thực tiễn** của nghiên cứu, với tầm nhìn ứng dụng trong nhiều lĩnh vực quan trọng:

* **Giao tiếp người-máy**: Hệ thống nhận diện cảm xúc có thể được tích hợp vào các trợ lý ảo, robot, hoặc chatbot, giúp chúng hiểu và phản hồi phù hợp với trạng thái cảm xúc của người dùng, từ đó nâng cao trải nghiệm giao tiếp.
* **Giám sát an ninh**: Công nghệ này có thể được sử dụng trong các hệ thống camera giám sát để phát hiện các hành vi bất thường hoặc trạng thái cảm xúc đáng ngờ, hỗ trợ đảm bảo an toàn công cộng.
* **Hỗ trợ tâm lý**: Trong y tế, hệ thống nhận diện cảm xúc có thể hỗ trợ các chuyên gia tâm lý theo dõi trạng thái cảm xúc của bệnh nhân, đặc biệt trong các trường hợp rối loạn tâm thần hoặc trầm cảm.
* **Giáo dục và thương mại**: Việc nhận diện cảm xúc có thể được áp dụng để cá nhân hóa trải nghiệm học tập hoặc phân tích phản hồi của khách hàng trong các chiến dịch tiếp thị.

Bằng cách kết nối các mục tiêu kỹ thuật với các ứng dụng thực tiễn, luận văn không chỉ giới hạn trong phạm vi học thuật mà còn mở ra tiềm năng triển khai trong đời sống, đặc biệt tại Việt Nam, nơi các công nghệ AI còn đang trong giai đoạn phát triển.

Bốn mục tiêu trên không hoạt động độc lập mà được thiết kế để hỗ trợ lẫn nhau, tạo nên một chu trình nghiên cứu hoàn chỉnh. Việc phát triển Residual Masking Network cung cấp nền tảng kỹ thuật cốt lõi, trong khi tập dữ liệu VEMO đảm bảo tính phù hợp với bối cảnh địa phương. Phương pháp học kết hợp nâng cao hiệu suất của mô hình, và quá trình đánh giá chặt chẽ khẳng định giá trị khoa học của công trình. Sự nhấn mạnh vào ứng dụng thực tiễn là sợi dây kết nối, đảm bảo rằng các thành tựu kỹ thuật không chỉ dừng lại ở lý thuyết mà còn mang lại lợi ích cụ thể cho xã hội.

Tầm nhìn của luận văn còn thể hiện qua việc tác giả không chỉ tập trung vào việc giải quyết bài toán nhận diện cảm xúc mà còn đặt mục tiêu xây dựng một nền tảng lâu dài cho nghiên cứu AI tại Việt Nam. Việc công khai tập dữ liệu VEMO và các kết quả thí nghiệm là minh chứng cho cam kết này, tạo điều kiện cho các nhà nghiên cứu khác tiếp tục phát triển và mở rộng bài toán trong tương lai.

# **Cơ sở lý thuyết**

Luận văn **“Nhận diện cảm xúc mặt người sử dụng mạng học sâu có cơ chế chú ý”** được xây dựng trên một nền tảng lý thuyết vững chắc, kết hợp các khái niệm cốt lõi từ tâm lý học, khoa học máy tính, và thị giác máy tính. Những cơ sở lý thuyết này không chỉ cung cấp căn cứ khoa học cho việc triển khai nghiên cứu mà còn định hướng rõ ràng cho quá trình thiết kế mô hình, thu thập dữ liệu, và đánh giá kết quả. Dưới đây là phân tích chi tiết và sâu hơn về các nền tảng lý thuyết chính được sử dụng trong luận văn, bao gồm nhận diện cảm xúc khuôn mặt, mạng nơ-ron tích chập, cơ chế chú ý, tập dữ liệu chuẩn, và các phương pháp trực quan hóa.

## **Nhận diện cảm xúc khuôn mặt**

Cơ sở lý thuyết đầu tiên và quan trọng nhất của luận văn là lĩnh vực **nhận diện cảm xúc khuôn mặt**, dựa trên các nghiên cứu tiên phong của nhà tâm lý học **Paul Ekman**. Ekman đã xác định **sáu biểu cảm cảm xúc cơ bản** – **giận dữ, ghê tởm, sợ hãi, hạnh phúc, buồn bã, và ngạc nhiên** – được coi là phổ quát trong mọi nền văn hóa. Các biểu cảm này không chỉ là sản phẩm của cảm xúc nội tại mà còn được thể hiện thông qua các chuyển động cơ mặt cụ thể, có thể quan sát và phân tích một cách hệ thống. Công trình của Ekman đã đặt nền móng cho việc hiểu cách con người giao tiếp phi ngôn ngữ thông qua khuôn mặt, mở ra cánh cửa cho các ứng dụng công nghệ trong việc mô phỏng và diễn giải các tín hiệu này.

Để mã hóa và phân tích các chuyển động trên khuôn mặt, luận văn sử dụng **Hệ thống mã hóa cơ mặt (Facial Action Coding System - FACS)**, cũng do Ekman và cộng sự phát triển. FACS chia nhỏ các chuyển động khuôn mặt thành các **đơn vị hành động (Action Units - AUs)**, mỗi đơn vị đại diện cho một cơ mặt hoặc nhóm cơ hoạt động cụ thể (ví dụ, nhíu mày, nâng khóe miệng). Hệ thống này cung cấp một ngôn ngữ chuẩn để mô tả và phân loại các biểu cảm, từ đó hỗ trợ việc xây dựng các mô hình máy học nhằm nhận diện cảm xúc. Trong luận văn, FACS không được sử dụng trực tiếp để trích xuất đặc trưng thủ công mà đóng vai trò là nền tảng lý thuyết, giúp định hình cách các mô hình học sâu tiếp cận bài toán phân loại cảm xúc. Việc hiểu rõ các biểu cảm cơ bản và cách chúng được mã hóa đã giúp tác giả xác định phạm vi của bài toán, bao gồm cả việc mở rộng sang trạng thái trung lập (neutral) trong các tập dữ liệu như FER2013.

Lý thuyết về cảm xúc khuôn mặt của Ekman cũng đặt ra các thách thức mà luận văn cần giải quyết, chẳng hạn như sự khác biệt về văn hóa trong cách thể hiện cảm xúc, các biểu cảm hỗn hợp (mixed emotions), và ảnh hưởng của các yếu tố ngoại cảnh như ánh sáng hay góc chụp. Những yếu tố này đã định hướng cho việc thiết kế mô hình và thu thập dữ liệu, đặc biệt là tập dữ liệu VEMO, nhằm phản ánh đặc điểm cảm xúc của người Việt Nam.

## **Mạng nơ-ron tích chập (CNN)**

**Mạng nơ-ron tích chập (Convolutional Neural Network - CNN)** là trụ cột kỹ thuật thứ hai của luận văn, đóng vai trò nền tảng cho việc trích xuất đặc trưng từ hình ảnh khuôn mặt. CNN là một loại mạng nơ-ron nhân tạo được thiết kế đặc biệt để xử lý dữ liệu có cấu trúc lưới, chẳng hạn như hình ảnh, nhờ vào các lớp tích chập (convolutional layers), lớp gộp (pooling layers), và lớp kết nối đầy đủ (fully connected layers). Các lớp tích chập sử dụng các bộ lọc (filters) để quét qua hình ảnh, tự động học và trích xuất các đặc trưng ở nhiều cấp độ, từ các đặc trưng đơn giản (như cạnh, góc) đến các đặc trưng phức tạp (như mắt, miệng, hoặc hình dạng khuôn mặt).

Luận văn tham khảo một loạt các kiến trúc CNN hiện đại, đã được chứng minh hiệu quả trong bài toán phân loại hình ảnh, bao gồm:

* **ResNet (Residual Network)**: Với các kết nối tắt (skip connections), ResNet giúp giải quyết vấn đề biến mất đạo hàm (vanishing gradient), cho phép huấn luyện các mạng sâu hơn mà vẫn duy trì hiệu suất cao. ResNet18, ResNet50, và ResNet152 được đề cập trong luận văn là các ví dụ điển hình.
* **VGG**: Được phát triển bởi Đại học Oxford, VGG sử dụng các lớp tích chập nhỏ (3x3) xếp chồng lên nhau để tạo ra mạng sâu, với VGG19 là một phiên bản được sử dụng trong nghiên cứu.
* **Inception (GoogleNet)**: Kiến trúc này tối ưu hóa hiệu suất bằng cách sử dụng các mô-đun Inception, kết hợp nhiều phép tích chập với kích thước khác nhau trong cùng một lớp.
* **DenseNet**: Với các kết nối dày đặc (dense connections), DenseNet tái sử dụng đặc trưng từ các lớp trước, giảm số lượng tham số và tăng hiệu quả tính toán.

Những kiến trúc này không chỉ cung cấp cơ sở để so sánh hiệu suất với mô hình đề xuất của luận văn mà còn truyền cảm hứng cho việc thiết kế **Residual Masking Network**. CNN được sử dụng trong luận văn để tự động trích xuất các đặc trưng từ hình ảnh khuôn mặt mà không cần dựa vào các phương pháp thủ công như trích xuất đặc trưng dựa trên FACS hay SIFT, từ đó tăng tính linh hoạt và khả năng xử lý dữ liệu phức tạp. Lý thuyết về CNN cũng định hướng cho việc chuẩn hóa dữ liệu (ví dụ, đưa hình ảnh về kích thước 224x224x3) và thiết lập các tham số huấn luyện như tốc độ học, momentum, và batch size.

## **Cơ chế chú ý (Attention Mechanism)**

**Cơ chế chú ý (Attention Mechanism)** là một thành phần lý thuyết đột phá, được tích hợp vào mô hình chính của luận văn – **Residual Masking Network** – để tối ưu hóa việc trích xuất đặc trưng. Lấy cảm hứng từ cách con người tập trung vào các chi tiết quan trọng khi quan sát, cơ chế chú ý cho phép mô hình học sâu ưu tiên các vùng nhất định trên hình ảnh (ví dụ, mắt, miệng, hoặc lông mày) thay vì xử lý toàn bộ hình ảnh một cách đồng đều. Điều này đặc biệt quan trọng trong bài toán nhận diện cảm xúc, bởi các vùng trên khuôn mặt mang thông tin cảm xúc khác nhau về mức độ và ý nghĩa.

Trong luận văn, cơ chế chú ý được triển khai thông qua kỹ thuật “masking” (che phủ), giúp mạng học cách tạo ra các trọng số (weights) để nhấn mạnh hoặc bỏ qua các vùng cụ thể trên khuôn mặt. Residual Masking Network kết hợp các kết nối tắt của ResNet với cơ chế chú ý, tạo ra một kiến trúc mạnh mẽ có khả năng học các đặc trưng phân biệt cao cấp (high-level features) liên quan đến cảm xúc. Lý thuyết về cơ chế chú ý được củng cố bởi các nghiên cứu gần đây trong học sâu, chẳng hạn như các mô hình Transformer và các mô-đun chú ý như CBAM (Convolutional Block Attention Module) hay BAM (Bottleneck Attention Module), được luận văn tham khảo để thiết kế mô hình.

Cơ chế chú ý không chỉ cải thiện độ chính xác mà còn tăng tính giải thích của mô hình, bởi nó cung cấp thông tin về việc mô hình tập trung vào đâu khi đưa ra dự đoán. Điều này đặc biệt hữu ích khi kết hợp với các phương pháp trực quan hóa như Grad-CAM, giúp làm rõ cách Residual Masking Network xử lý dữ liệu và ra quyết định.

## **Tập dữ liệu chuẩn**

Tập dữ liệu đóng vai trò quan trọng trong việc cung cấp cơ sở thực nghiệm để huấn luyện và đánh giá mô hình. Luận văn sử dụng hai tập dữ liệu chính, mỗi cái mang một ý nghĩa lý thuyết riêng:

* **Tập dữ liệu FER2013**: Đây là một tập dữ liệu công khai phổ biến trong lĩnh vực nhận diện cảm xúc, chứa hơn 35,000 hình ảnh khuôn mặt với độ phân giải thấp (48x48 pixel). FER2013 bao gồm bảy loại cảm xúc: giận dữ, ghê tởm, sợ hãi, hạnh phúc, buồn bã, ngạc nhiên, và trung lập. Tập dữ liệu này được thiết kế để mô phỏng các điều kiện thực tế, với hình ảnh thu thập từ internet, dẫn đến sự đa dạng về ánh sáng, góc chụp, và chất lượng. Lý thuyết đằng sau FER2013 nhấn mạnh tính thách thức của bài toán nhận diện cảm xúc trong môi trường không lý tưởng, từ đó cung cấp một chuẩn mực để đánh giá khả năng tổng quát hóa của mô hình.
* **Tập dữ liệu VEMO**: Do chính tác giả xây dựng, VEMO tập trung vào đặc điểm khuôn mặt của người Việt Nam, phản ánh các sắc thái cảm xúc và đặc điểm văn hóa riêng biệt. Việc tạo ra VEMO dựa trên lý thuyết rằng các biểu cảm cảm xúc có thể khác nhau giữa các nền văn hóa, như nghiên cứu của Rachael E. Jack và cộng sự đã chỉ ra. VEMO không chỉ là một nguồn dữ liệu mà còn là một đóng góp lý thuyết, bổ sung vào kho tài nguyên còn hạn chế cho các nghiên cứu nhận diện cảm xúc tại Việt Nam.

Cả hai tập dữ liệu đều được sử dụng để đảm bảo rằng mô hình không chỉ hiệu quả trên dữ liệu quốc tế mà còn phù hợp với bối cảnh địa phương, từ đó củng cố tính ứng dụng của nghiên cứu.

## **Các phương pháp trực quan hóa**

Để tăng tính minh bạch và giải thích được cách mô hình đưa ra quyết định, luận văn sử dụng hai phương pháp trực quan hóa chính:

* **Gradient-weighted Class Activation Mapping (Grad-CAM)**: Grad-CAM là một kỹ thuật trực quan hóa nổi tiếng, sử dụng đạo hàm riêng của đầu ra dự đoán đối với các đặc trưng trong lớp tích chập cuối cùng để tạo ra bản đồ nhiệt (heatmap). Bản đồ này chỉ ra các vùng trên hình ảnh mà mô hình tập trung khi phân loại cảm xúc, chẳng hạn như mắt hoặc miệng. Grad-CAM giúp giải thích tại sao mô hình đưa ra một dự đoán cụ thể, từ đó tăng độ tin cậy và khả năng debug của hệ thống.
* **Phương pháp gộp trung bình đặc trưng**: Kỹ thuật này lấy trung bình các đặc trưng từ các lớp tích chập theo chiều kênh, cung cấp một góc nhìn tổng quan về cách mô hình xử lý dữ liệu. Phương pháp này bổ sung cho Grad-CAM, giúp phân tích sâu hơn các đặc trưng mà mô hình học được.

Lý thuyết đằng sau các phương pháp trực quan hóa này dựa trên nhu cầu giải thích các mô hình học sâu, vốn thường được coi là “hộp đen” do tính phức tạp của chúng. Bằng cách sử dụng Grad-CAM và gộp trung bình đặc trưng, luận văn không chỉ đánh giá hiệu suất mà còn cung cấp bằng chứng trực quan về cách Residual Masking Network hoạt động, từ đó củng cố tính khoa học và thuyết phục của nghiên cứu.

Các nền tảng lý thuyết trên không hoạt động độc lập mà được tích hợp chặt chẽ để hỗ trợ bài toán nhận diện cảm xúc. Lý thuyết về cảm xúc khuôn mặt của Ekman và FACS cung cấp cơ sở để xác định các lớp cảm xúc và hiểu cách chúng được thể hiện. CNN và các kiến trúc hiện đại như ResNet, VGG đặt nền tảng kỹ thuật cho việc trích xuất đặc trưng. Cơ chế chú ý nâng cao khả năng phân tích của mô hình, trong khi các tập dữ liệu FER2013 và VEMO đảm bảo tính thực tiễn và phù hợp với bối cảnh. Cuối cùng, các phương pháp trực quan hóa tăng tính minh bạch, giúp kết nối lý thuyết với thực nghiệm.

Những nền tảng lý thuyết này đã định hướng rõ ràng cho việc thiết kế **Residual Masking Network**, lựa chọn tập dữ liệu, và triển khai các thí nghiệm. Chúng cũng phản ánh sự hiểu biết sâu sắc của tác giả về cả khía cạnh tâm lý học (cách con người thể hiện cảm xúc) và kỹ thuật (cách máy tính mô phỏng quá trình này), từ đó tạo nên một công trình nghiên cứu toàn diện và có giá trị.

# **Phương pháp thu thập và xử lý thông tin**

Phương pháp nghiên cứu trong luận văn **“Nhận diện cảm xúc mặt người sử dụng mạng học sâu có cơ chế chú ý”** được triển khai một cách có hệ thống, khoa học và chi tiết, nhằm đảm bảo tính chặt chẽ, công bằng và khả năng tái lập của các thí nghiệm. Phương pháp này bao gồm các giai đoạn chính: **thu thập dữ liệu**, **xử lý dữ liệu**, **thực nghiệm**, và **trực quan hóa kết quả**. Mỗi giai đoạn được thiết kế cẩn thận để hỗ trợ mục tiêu phát triển một hệ thống nhận diện cảm xúc khuôn mặt hiệu quả, tận dụng các kỹ thuật học sâu tiên tiến và phù hợp với bối cảnh Việt Nam. Dưới đây là phân tích chi tiết và sâu hơn về các phương pháp này, làm rõ cách chúng đóng góp vào thành công của nghiên cứu.

## **Thu thập dữ liệu**

Việc thu thập dữ liệu là bước đầu tiên và quan trọng, cung cấp nguyên liệu để huấn luyện và đánh giá các mô hình học sâu. Luận văn sử dụng hai tập dữ liệu chính, mỗi cái mang đặc điểm riêng biệt và phục vụ các mục tiêu khác nhau:

* **Tập dữ liệu FER2013**: Đây là một tập dữ liệu công khai phổ biến trong lĩnh vực nhận diện cảm xúc, được giới thiệu tại cuộc thi ICML 2013. FER2013 chứa hơn **35,000 hình ảnh khuôn mặt** với độ phân giải thấp (**48x48 pixel**, đơn sắc), được thu thập tự động từ internet thông qua các công cụ tìm kiếm hình ảnh. Tập dữ liệu này bao gồm **bảy loại cảm xúc**: giận dữ, ghê tởm, sợ hãi, hạnh phúc, buồn bã, ngạc nhiên, và trung lập. FER2013 được chia thành ba phần: tập huấn luyện (28,709 ảnh), tập validation (3,589 ảnh), và tập kiểm tra (3,589 ảnh).  
    
   Đặc điểm nổi bật của FER2013 là sự đa dạng về điều kiện thực tế, bao gồm các yếu tố như ánh sáng không đồng đều, góc chụp khác nhau, và chất lượng hình ảnh thấp. Tuy nhiên, tập dữ liệu này cũng có hạn chế, như sự mất cân bằng giữa các lớp cảm xúc (ví dụ, cảm xúc hạnh phúc chiếm tỷ lệ lớn hơn so với ghê tởm) và sự hiện diện của nhiễu (hình ảnh không phải khuôn mặt hoặc gắn nhãn sai). Trong luận văn, FER2013 được chọn làm chuẩn mực quốc tế, cho phép tác giả đánh giá hiệu quả của mô hình đề xuất so với các nghiên cứu toàn cầu, đồng thời kiểm tra khả năng tổng quát hóa trên dữ liệu phức tạp.
* **Tập dữ liệu VEMO**: Đây là tập dữ liệu do chính tác giả xây dựng, tập trung vào **khuôn mặt người Việt Nam**. VEMO được thiết kế để phản ánh các đặc điểm văn hóa, sắc thái cảm xúc, và sự đa dạng về biểu cảm của người Việt Nam, vốn không được đại diện đầy đủ trong các tập dữ liệu quốc tế như FER2013. Việc xây dựng VEMO là một đóng góp quan trọng của luận văn, khắc phục sự thiếu hụt dữ liệu địa phương trong nghiên cứu nhận diện cảm xúc tại Việt Nam.  
    
   Quá trình thu thập VEMO đòi hỏi sự tỉ mỉ trong việc chụp ảnh, ghi lại video, hoặc sử dụng các nguồn hình ảnh thực tế, sau đó gắn nhãn thủ công hoặc bán tự động để đảm bảo tính chính xác. Mặc dù luận văn không nêu rõ số lượng hình ảnh cụ thể trong VEMO, tập dữ liệu này được chuẩn hóa để phù hợp với bài toán, bao gồm các cảm xúc tương tự FER2013. VEMO không chỉ cung cấp cơ sở để đánh giá mô hình trong bối cảnh Việt Nam mà còn mở ra tiềm năng cho các nghiên cứu tiếp theo, đặc biệt trong các ứng dụng thực tiễn như giao tiếp người-máy hoặc hỗ trợ tâm lý tại địa phương.

Việc sử dụng cả FER2013 và VEMO thể hiện tầm nhìn toàn diện của tác giả: vừa đảm bảo khả năng cạnh tranh trên phạm vi quốc tế, vừa đáp ứng nhu cầu nghiên cứu và ứng dụng tại Việt Nam. Hai tập dữ liệu này bổ trợ lẫn nhau, với FER2013 cung cấp khối lượng dữ liệu lớn và VEMO mang tính đặc thù văn hóa.

## **Xử lý dữ liệu**

Xử lý dữ liệu là bước quan trọng để đảm bảo chất lượng đầu vào cho các mô hình học sâu, giúp tối ưu hóa hiệu suất và giảm thiểu các vấn đề như quá khớp (overfitting). Luận văn đã triển khai một quy trình xử lý dữ liệu cẩn thận, bao gồm các kỹ thuật sau:

* **Chuẩn hóa kích thước hình ảnh**: Hình ảnh trong cả FER2013 và VEMO được chuẩn hóa về kích thước **224x224x3** (chiều cao x chiều rộng x số kênh màu RGB). Đây là kích thước phổ biến cho các kiến trúc CNN hiện đại như ResNet, VGG, hay Inception, đảm bảo khả năng tương thích với các mô hình được sử dụng trong nghiên cứu. Đối với FER2013, việc nâng cấp từ độ phân giải 48x48 pixel lên 224x224 yêu cầu các kỹ thuật nội suy (interpolation), chẳng hạn như nội suy song tuyến (bilinear interpolation), để bảo toàn thông tin hình ảnh.
* **Tăng cường dữ liệu (Data Augmentation)**: Để tăng tính đa dạng của dữ liệu và giảm nguy cơ quá khớp, tác giả áp dụng một loạt kỹ thuật tăng cường dữ liệu, bao gồm:
* **Xoay (rotation)**: Xoay hình ảnh trong một phạm vi góc nhỏ để mô phỏng các góc chụp khác nhau.
* **Cắt (cropping)**: Cắt ngẫu nhiên một phần hình ảnh để tạo ra các góc nhìn mới.
* **Thay đổi độ sáng (brightness adjustment)**: Điều chỉnh độ sáng để mô phỏng các điều kiện ánh sáng khác nhau.
* **Lật ngang (horizontal flipping)**: Lật hình ảnh theo trục ngang để tăng tính đối xứng.
* Những kỹ thuật này giúp mô hình học được các đặc trưng bất biến (invariant features) đối với các biến đổi hình học và môi trường, từ đó cải thiện khả năng tổng quát hóa trên dữ liệu mới.
* **Tiền xử lý bằng OpenCV**: Tác giả sử dụng thư viện **OpenCV**, một công cụ mạnh mẽ trong xử lý hình ảnh, để làm sạch và chuẩn hóa dữ liệu. Các bước tiền xử lý có thể bao gồm:
* **Xóa nhiễu (noise removal)**: Áp dụng các bộ lọc như Gaussian blur để loại bỏ nhiễu từ hình ảnh chất lượng thấp.
* **Cân bằng sáng (histogram equalization)**: Cải thiện độ tương phản của hình ảnh để làm nổi bật các đặc trưng khuôn mặt.
* **Phát hiện khuôn mặt (face detection)**: Sử dụng các thuật toán như Haar Cascade hoặc DLib để cắt bỏ các vùng không liên quan, đảm bảo hình ảnh chỉ chứa khuôn mặt.

Quá trình xử lý dữ liệu này được thiết kế để tối ưu hóa đầu vào cho các mô hình học sâu, đồng thời đảm bảo rằng dữ liệu từ cả FER2013 và VEMO có thể được sử dụng một cách đồng nhất trong các thí nghiệm.

## **Phương pháp thực nghiệm**

Phương pháp thực nghiệm là cốt lõi của luận văn, nơi các ý tưởng lý thuyết được hiện thực hóa thông qua thiết kế mô hình, huấn luyện, và đánh giá. Tác giả đã triển khai một quy trình thực nghiệm có hệ thống, bao gồm các thành phần sau:

* **Residual Masking Network**: Đây là mô hình chính của luận văn, được thiết kế đặc biệt để tích hợp **cơ chế chú ý (attention mechanism)** vào quá trình trích xuất đặc trưng. Residual Masking Network lấy cảm hứng từ ResNet với các kết nối tắt (skip connections), nhưng bổ sung các lớp “masking” để ưu tiên các vùng quan trọng trên khuôn mặt (như mắt, miệng, lông mày). Mô hình này có số lượng tham số lớn – **142.9 triệu tham số** – cho thấy độ phức tạp cao, nhưng cũng mang lại khả năng học các đặc trưng phân biệt mạnh mẽ. Cơ chế chú ý trong Residual Masking Network hoạt động bằng cách tạo ra các bản đồ trọng số (attention maps), giúp mô hình tập trung vào các vùng có liên quan đến cảm xúc thay vì toàn bộ hình ảnh. Điều này đặc biệt hiệu quả trong bài toán nhận diện cảm xúc, nơi các đặc trưng cục bộ (local features) đóng vai trò quan trọng hơn so với bối cảnh toàn cục (global context). Mô hình được thiết kế để tối ưu hóa hiệu suất trên cả FER2013 và VEMO, với sự cân nhắc đến sự khác biệt về độ phân giải và đặc điểm văn hóa.
* **Học kết hợp (Ensemble Learning)**: Để nâng cao độ chính xác, tác giả sử dụng phương pháp học kết hợp, kết hợp **Residual Masking Network** với **sáu mô hình CNN khác**, chẳng hạn như ResNet18, ResNet34, VGG19, DenseNet121, và InceptionV3. Học kết hợp tận dụng thế mạnh của từng mô hình riêng lẻ, chẳng hạn như khả năng học sâu của ResNet, độ chính xác của VGG, hay hiệu quả tính toán của DenseNet, để tạo ra một hệ thống mạnh mẽ hơn. Quá trình kết hợp được thực hiện bằng cách lấy trung bình trọng số (weighted averaging) hoặc bỏ phiếu đa số (majority voting) dựa trên dự đoán của các mô hình. Phương pháp này giúp giảm thiểu sai số và cải thiện khả năng tổng quát hóa, đặc biệt trên các tập dữ liệu có tính đa dạng cao như FER2013 và VEMO.
* **Cấu hình huấn luyện**: Các mô hình được huấn luyện với các tham số được tối ưu hóa cẩn thận:
* **Tốc độ học (learning rate)**: 0.0001, đảm bảo mô hình học một cách từ từ và ổn định.
* **Momentum**: 0.9, giúp tăng tốc quá trình hội tụ bằng cách sử dụng động lượng từ các bước cập nhật trước.
* **Weight decay**: 0.001, ngăn chặn hiện tượng quá khớp bằng cách áp dụng chính quy hóa L2.
* **Batch size**: 48, cân bằng giữa tốc độ huấn luyện và độ chính xác.
* **Số epoch tối đa**: 50, với cơ chế **dừng sớm (early stopping)** nếu độ chính xác trên tập validation không cải thiện sau 8 epoch liên tiếp.
* Quá trình huấn luyện sử dụng thuật toán tối ưu hóa **Stochastic Gradient Descent (SGD)** hoặc các biến thể như Adam, kết hợp với hàm mất mát **cross-entropy loss**, phù hợp với bài toán phân loại đa lớp. Các tham số này được điều chỉnh dựa trên các thử nghiệm ban đầu để đảm bảo hiệu suất tối ưu.
* **Đánh giá mô hình**: Hiệu quả của các mô hình được đánh giá thông qua hai chỉ số chính:
* **Độ chính xác (accuracy)**: Tỷ lệ dự đoán đúng trên tập kiểm tra, là thước đo trực quan nhất về hiệu suất.
* **Ma trận bối rối (confusion matrix)**: Cung cấp cái nhìn chi tiết về hiệu suất của mô hình trên từng lớp cảm xúc, giúp xác định các lớp dễ bị nhầm lẫn (ví dụ, giận dữ và buồn bã).
* Các mô hình đề xuất được so sánh với các kiến trúc hiện đại như VGG19, ResNet18, ResNet34, DenseNet121, và InceptionV3, cũng như các phương pháp được báo cáo trong tài liệu khoa học (chẳng hạn như Deep-Emotion, Ensemble MLCNNs). Quá trình so sánh này đảm bảo rằng Residual Masking Network và phương pháp học kết hợp có tính cạnh tranh trên phạm vi quốc tế.

## **Trực quan hóa kết quả**

Để tăng tính giải thích và minh bạch, luận văn sử dụng các phương pháp trực quan hóa nhằm làm rõ cách mô hình đưa ra quyết định:

* **Gradient-weighted Class Activation Mapping (Grad-CAM)**: Grad-CAM tạo ra **bản đồ nhiệt (heatmap)**, chỉ ra các vùng trên hình ảnh khuôn mặt mà mô hình tập trung khi dự đoán cảm xúc. Ví dụ, bản đồ nhiệt có thể làm nổi bật vùng mắt khi dự đoán cảm xúc sợ hãi, hoặc vùng miệng khi dự đoán hạnh phúc. Kỹ thuật này giúp xác nhận rằng Residual Masking Network thực sự ưu tiên các vùng quan trọng, đồng thời cung cấp công cụ để debug và cải thiện mô hình.
* **Phương pháp gộp trung bình đặc trưng**: Phương pháp này lấy trung bình các đặc trưng từ các lớp tích chập theo chiều kênh, tạo ra một biểu diễn trực quan về cách mô hình xử lý dữ liệu. Kết quả gộp trung bình giúp phân tích các đặc trưng cấp cao (high-level features) mà mô hình học được, chẳng hạn như các mẫu liên quan đến nếp nhăn hoặc hình dạng miệng.

Những phương pháp trực quan hóa này không chỉ làm tăng tính thuyết phục của nghiên cứu mà còn đóng vai trò như một cầu nối giữa lý thuyết và thực tiễn, giúp các nhà nghiên cứu và người dùng cuối hiểu rõ hơn về hoạt động của mô hình.

Phương pháp thu thập và xử lý thông tin được thiết kế để đảm bảo **tính công bằng** và **khả năng tái lập**:

* **Tính công bằng**: Việc sử dụng cả FER2013 (chuẩn quốc tế) và VEMO (đặc thù Việt Nam) đảm bảo rằng mô hình được đánh giá trong các bối cảnh đa dạng. Các kỹ thuật tăng cường dữ liệu và chuẩn hóa giúp giảm thiểu thiên kiến từ dữ liệu đầu vào.
* **Khả năng tái lập**: Các tham số huấn luyện (tốc độ học, batch size, v.v.), cấu hình mô hình, và quy trình xử lý dữ liệu được mô tả chi tiết, cho phép các nhà nghiên cứu khác tái tạo thí nghiệm. Việc sử dụng các thư viện phổ biến như OpenCV, PyTorch, hoặc TensorFlow cũng tăng tính truy cập của nghiên cứu.

Phương pháp thu thập và xử lý thông tin đã góp phần quan trọng vào thành công của luận văn:

* **Dữ liệu đa dạng**: Sự kết hợp giữa FER2013 và VEMO đảm bảo rằng mô hình được huấn luyện trên cả dữ liệu quốc tế và địa phương, tăng tính ứng dụng thực tiễn.
* **Xử lý cẩn thận**: Các kỹ thuật chuẩn hóa và tăng cường dữ liệu giúp mô hình học được các đặc trưng mạnh mẽ, bất biến với các điều kiện thực tế.
* **Thực nghiệm tiên tiến**: Residual Masking Network và học kết hợp thể hiện sự đổi mới kỹ thuật, trong khi cấu hình huấn luyện và đánh giá chặt chẽ đảm bảo độ tin cậy của kết quả.
* **Trực quan hóa thuyết phục**: Grad-CAM và gộp trung bình đặc trưng cung cấp cái nhìn sâu sắc về hoạt động của mô hình, tăng tính minh bạch và giá trị khoa học.

# **Kết quả đạt được**

Luận văn **“Nhận diện cảm xúc mặt người sử dụng mạng học sâu có cơ chế chú ý”** đã đạt được những thành tựu đáng chú ý, khẳng định tính hiệu quả của các phương pháp đề xuất và mang lại giá trị khoa học cũng như thực tiễn. Các kết quả này được thể hiện qua hiệu suất của mô hình **Residual Masking Network** trên các tập dữ liệu **FER2013** và **VEMO**, đóng góp của tập dữ liệu VEMO, khả năng trực quan hóa và giải thích mô hình, cũng như tiềm năng ứng dụng thực tế. Dưới đây là phân tích chi tiết và sâu hơn về những kết quả này, làm rõ ý nghĩa và tác động của chúng trong bối cảnh nghiên cứu nhận diện cảm xúc khuôn mặt.

## **Hiệu quả của Residual Masking Network**

**Residual Masking Network**, mô hình trung tâm của luận văn, đã chứng minh được hiệu suất vượt trội trong bài toán phân loại cảm xúc khuôn mặt, nhờ vào việc tích hợp **cơ chế chú ý** để tập trung vào các vùng quan trọng như mắt, miệng, và lông mày. Kết quả cụ thể bao gồm:

* **Trên tập dữ liệu FER2013**: Residual Masking Network đạt **độ chính xác 74.14%** trên tập kiểm tra, vượt qua nhiều kiến trúc CNN hiện đại được sử dụng làm cơ sở so sánh:
* **VGG19**: 70.8%, một mô hình sâu với số lượng tham số lớn nhưng hiệu suất thấp hơn do thiếu cơ chế chú ý.
* **ResNet18**: 72.9%, hiệu quả hơn VGG19 nhờ kết nối tắt nhưng vẫn không bằng Residual Masking Network.
* **DenseNet121**: 73.16%, với kết nối dày đặc giúp giảm tham số nhưng vẫn thua kém về độ chính xác.
* Độ chính xác 74.14% trên FER2013 – một tập dữ liệu đầy thách thức với hình ảnh độ phân giải thấp, điều kiện ánh sáng đa dạng, và sự mất cân bằng giữa các lớp cảm xúc – cho thấy Residual Masking Network không chỉ mạnh mẽ mà còn có khả năng học được các đặc trưng phân biệt cao cấp (high-level features). Cơ chế chú ý đã giúp mô hình tập trung vào các vùng mang thông tin cảm xúc quan trọng, từ đó cải thiện khả năng phân loại so với các kiến trúc truyền thống.
* **Học kết hợp (Ensemble Learning)**: Để nâng cao hiệu suất, tác giả đã kết hợp Residual Masking Network với **sáu mô hình CNN khác** (chẳng hạn như ResNet18, ResNet34, VGG19, DenseNet121, InceptionV3, và một mô hình tùy chỉnh khác). Phương pháp học kết hợp này đạt **độ chính xác cao nhất 76.82%** trên FER2013, vượt qua các phương pháp tiên tiến được báo cáo trong tài liệu khoa học:
* **Ensemble 8 CNNs** (Pramerdorfer và Kampel, 2016): 75.2%, một phương pháp học kết hợp phức tạp nhưng kém hơn về độ chính xác.
* **CNNs với BOVW (Bag of Visual Words)** (Ionescu và cộng sự, 2013): 75.42%, một cách tiếp cận kết hợp học sâu và đặc trưng thủ công nhưng vẫn không đạt hiệu quả bằng.
* Kết quả 76.82% là một bước tiến đáng kể, đặt phương pháp của luận văn vào nhóm các giải pháp hàng đầu cho bài toán nhận diện cảm xúc trên FER2013. Học kết hợp tận dụng thế mạnh của nhiều mô hình, giúp giảm thiểu sai số và tăng tính ổn định của dự đoán, đặc biệt trong các trường hợp dữ liệu phức tạp hoặc nhiễu. Điều này cũng phản ánh sự sáng tạo của tác giả trong việc tối ưu hóa hiệu suất bằng cách kết hợp các kỹ thuật tiên tiến.

## **Hiệu quả trên tập dữ liệu VEMO**

Tập dữ liệu **VEMO**, do tác giả tự xây dựng để đại diện cho khuôn mặt người Việt Nam, là một bối cảnh thử nghiệm quan trọng để đánh giá khả năng tổng quát hóa của mô hình trên dữ liệu đặc thù văn hóa. Kết quả trên VEMO bao gồm:

* **Residual Masking Network** đạt **độ chính xác 65.949%**, cao hơn các mô hình so sánh:
* **ResNet18**: 63.94%, hiệu quả nhưng kém hơn do thiếu cơ chế chú ý.
* **ResNet34**: 64.84%, cải thiện so với ResNet18 nhưng vẫn không bằng Residual Masking Network.
* Mặc dù độ chính xác trên VEMO thấp hơn so với FER2013 (65.949% so với 74.14%), điều này có thể được giải thích bởi quy mô nhỏ hơn của VEMO, sự khác biệt về chất lượng hình ảnh, hoặc các sắc thái cảm xúc đặc thù của người Việt Nam. Tuy nhiên, việc Residual Masking Network vẫn vượt trội so với các mô hình khác cho thấy khả năng thích nghi của nó với dữ liệu mới, đặc biệt trong bối cảnh thiếu hụt tài nguyên dữ liệu địa phương. Kết quả này khẳng định rằng cơ chế chú ý không chỉ hiệu quả trên dữ liệu quốc tế mà còn phù hợp với các đặc điểm khuôn mặt và biểu cảm của người Việt Nam. Độ chính xác trên VEMO cũng mang ý nghĩa quan trọng vì nó chứng minh rằng mô hình có thể xử lý các bài toán nhận diện cảm xúc trong bối cảnh thực tế tại Việt Nam, nơi các ứng dụng như giám sát tâm lý, giáo dục, hoặc giao tiếp người-máy đang ngày càng được quan tâm. Kết quả này là minh chứng cho giá trị của việc xây dựng VEMO, mở ra tiềm năng cho các nghiên cứu tiếp theo trong lĩnh vực này.

## **Đóng góp của tập dữ liệu VEMO**

Việc xây dựng và công khai **tập dữ liệu VEMO** là một trong những đóng góp quan trọng nhất của luận văn, không chỉ về mặt kỹ thuật mà còn về ý nghĩa lâu dài cho cộng đồng nghiên cứu tại Việt Nam. VEMO được thiết kế để phản ánh **đặc điểm văn hóa và sắc thái cảm xúc của người Việt Nam**, khắc phục hạn chế của các tập dữ liệu quốc tế như FER2013, vốn chủ yếu bao gồm hình ảnh từ các quốc gia phương Tây.

Quá trình xây dựng VEMO đòi hỏi sự nỗ lực đáng kể trong việc thu thập hình ảnh, gắn nhãn cảm xúc, và chuẩn hóa dữ liệu để đảm bảo tính chính xác và đa dạng. Mặc dù luận văn không cung cấp thông tin chi tiết về quy mô cụ thể của VEMO, tập dữ liệu này được thiết kế để bao gồm các cảm xúc tương tự FER2013 (giận dữ, ghê tởm, sợ hãi, hạnh phúc, buồn bã, ngạc nhiên, trung lập), với sự tập trung vào các đặc trưng khuôn mặt và biểu cảm của người Việt Nam.

VEMO không chỉ đóng vai trò như một công cụ để đánh giá mô hình trong luận văn mà còn tạo nền tảng cho các nghiên cứu tiếp theo về nhận diện cảm xúc tại Việt Nam. Việc công khai tập dữ liệu này là một bước tiến chiến lược, khuyến khích các nhà nghiên cứu khác sử dụng và mở rộng VEMO, từ đó thúc đẩy sự phát triển của trí tuệ nhân tạo trong các ứng dụng mang tính địa phương. Đóng góp này đặc biệt ý nghĩa trong bối cảnh Việt Nam, nơi các tài nguyên dữ liệu cho AI còn hạn chế, và các bài toán liên quan đến cảm xúc con người vẫn đang ở giai đoạn sơ khai.

## **Trực quan hóa và giải thích mô hình**

Một trong những điểm mạnh của luận văn là khả năng giải thích cách mô hình đưa ra quyết định, thông qua các kỹ thuật trực quan hóa tiên tiến:

* **Bản đồ nhiệt từ Grad-CAM**: **Gradient-weighted Class Activation Mapping (Grad-CAM)** được sử dụng để tạo ra các bản đồ nhiệt, làm nổi bật các vùng trên khuôn mặt mà Residual Masking Network tập trung khi dự đoán cảm xúc. Kết quả cho thấy mô hình tập trung chính xác vào các vùng quan trọng, chẳng hạn như:
* **Vùng mắt**: Thường liên quan đến cảm xúc như sợ hãi hoặc buồn bã, nơi các chuyển động như mở to mắt hoặc nhíu mày mang thông tin quan trọng.
* **Vùng miệng**: Quan trọng trong các cảm xúc như hạnh phúc (nụ cười) hoặc giận dữ (mím môi).
* Các bản đồ nhiệt này không chỉ xác nhận hiệu quả của cơ chế chú ý trong Residual Masking Network mà còn tăng tính thuyết phục của nghiên cứu, bởi chúng cung cấp bằng chứng trực quan rằng mô hình thực sự học được các đặc trưng liên quan đến cảm xúc. Grad-CAM cũng giúp phát hiện các trường hợp dự đoán sai, hỗ trợ quá trình cải thiện mô hình trong tương lai.
* **Ma trận bối rối (Confusion Matrix)**: Ma trận bối rối cung cấp cái nhìn chi tiết về hiệu suất của mô hình trên từng lớp cảm xúc, đặc biệt trong việc xử lý dữ liệu mất cân bằng. Kết quả cho thấy Residual Masking Network xử lý tốt các cảm xúc phổ biến như **hạnh phúc** và **buồn bã**, vốn có số lượng mẫu lớn trong FER2013. Tuy nhiên, các cảm xúc ít mẫu hơn như **ghê tởm** hoặc **sợ hãi** có thể gặp khó khăn, dẫn đến một số nhầm lẫn giữa các lớp. Ma trận bối rối không chỉ đánh giá hiệu quả mà còn định hướng cho các cải tiến tiếp theo, chẳng hạn như tăng cường dữ liệu cho các lớp thiểu số.

Các kỹ thuật trực quan hóa này làm tăng tính minh bạch của mô hình, giúp giải quyết vấn đề “hộp đen” thường thấy trong học sâu. Chúng không chỉ có giá trị khoa học mà còn hỗ trợ các ứng dụng thực tiễn, nơi người dùng cần hiểu rõ lý do đằng sau các dự đoán của hệ thống.

## **Ứng dụng thực tế**

Luận văn không chỉ dừng lại ở các kết quả học thuật mà còn chứng minh tiềm năng ứng dụng thực tế của mô hình thông qua các thử nghiệm trên **hình ảnh thực tế**. Cụ thể, tác giả đã áp dụng Residual Masking Network trên các cảnh trong các bộ phim nổi tiếng như **“Mắt Biếc”** (một bộ phim Việt Nam) và **“Harry Potter”** (một bộ phim quốc tế). Những thử nghiệm này cho thấy mô hình có khả năng nhận diện cảm xúc chính xác trong các tình huống đời sống, với các điều kiện ánh sáng, góc chụp, và biểu cảm phức tạp hơn so với dữ liệu huấn luyện.

* Trong **“Mắt Biếc”**, mô hình đã nhận diện thành công các cảm xúc như buồn bã và hạnh phúc trong các cảnh tình cảm, phản ánh sự phù hợp với sắc thái cảm xúc của người Việt Nam.
* Trong **“Harry Potter”**, mô hình xử lý tốt các biểu cảm đa dạng trong các cảnh hành động hoặc kịch tính, cho thấy khả năng tổng quát hóa trên dữ liệu quốc tế.

Những thử nghiệm này không chỉ minh họa tính mạnh mẽ của mô hình mà còn làm nổi bật tiềm năng ứng dụng trong các lĩnh vực thực tiễn, bao gồm:

* **Giao tiếp người-máy**: Tích hợp vào các trợ lý ảo hoặc robot để hiểu và phản hồi theo cảm xúc của người dùng.
* **Giám sát an ninh**: Phát hiện các trạng thái cảm xúc bất thường trong hệ thống camera giám sát.
* **Hỗ trợ tâm lý**: Theo dõi cảm xúc của bệnh nhân trong các buổi trị liệu hoặc chẩn đoán tâm lý.
* **Giải trí và truyền thông**: Phân tích phản hồi cảm xúc của khán giả đối với nội dung phim, quảng cáo, hoặc trò chơi.

Việc thử nghiệm trên các hình ảnh thực tế là một bước tiến quan trọng, giúp kết nối lý thuyết với thực tiễn và khẳng định rằng mô hình không chỉ hiệu quả trong phòng thí nghiệm mà còn có thể triển khai trong các kịch bản đời sống.

Các kết quả đạt được của luận văn mang lại nhiều ý nghĩa và tác động đáng kể:

* **Về mặt khoa học**: Độ chính xác 76.82% trên FER2013 và 65.949% trên VEMO đặt nghiên cứu này vào nhóm các giải pháp hàng đầu trong lĩnh vực nhận diện cảm xúc. Residual Masking Network và phương pháp học kết hợp thể hiện sự đổi mới kỹ thuật, trong khi tập dữ liệu VEMO bổ sung một tài nguyên quý giá cho cộng đồng nghiên cứu.
* **Về mặt thực tiễn**: Các thử nghiệm trên hình ảnh thực tế và tiềm năng ứng dụng trong giao tiếp, an ninh, tâm lý, và giải trí cho thấy luận văn không chỉ mang tính học thuật mà còn có giá trị xã hội lớn, đặc biệt tại Việt Nam.
* **Về mặt văn hóa**: VEMO là một bước tiến trong việc đại diện cho người Việt Nam trong nghiên cứu AI, góp phần giảm sự phụ thuộc vào dữ liệu nước ngoài và tăng tính phù hợp của công nghệ với bối cảnh địa phương.

Những kết quả này không chỉ chứng minh tính hiệu quả của phương pháp đề xuất mà còn khẳng định khả năng của tác giả trong việc kết hợp lý thuyết, thực nghiệm, và ứng dụng thực tiễn để giải quyết một bài toán phức tạp.

# **Hạn chế của công trình**

Mặc dù luận văn **“Nhận diện cảm xúc mặt người sử dụng mạng học sâu có cơ chế chú ý”** đã đạt được nhiều thành tựu đáng kể, từ hiệu suất vượt trội của **Residual Masking Network** đến đóng góp quan trọng của tập dữ liệu **VEMO**, công trình vẫn tồn tại một số hạn chế. Những hạn chế này không làm giảm giá trị tổng thể của nghiên cứu mà thay vào đó phản ánh những thách thức vốn có trong lĩnh vực trí tuệ nhân tạo (AI) và học sâu, đồng thời mở ra các hướng phát triển tiềm năng cho các nghiên cứu tương lai. Dưới đây là phân tích chi tiết và sâu hơn về các hạn chế được nêu trong luận văn, cùng với bối cảnh và ý nghĩa của chúng.

## **Thời gian nghiên cứu hạn chế**

Một trong những hạn chế chính của luận văn là **thời gian thực hiện ngắn**, vốn giới hạn khả năng khám phá đầy đủ tiềm năng của các phương pháp đề xuất. Là một luận văn tốt nghiệp đại học, công trình được thực hiện trong một khung thời gian cố định, thường chỉ kéo dài vài tháng, điều này ảnh hưởng đến phạm vi thử nghiệm và tối ưu hóa mô hình.

* **Hạn chế trong thử nghiệm mô hình**: Do thời gian có hạn, tác giả chưa thể thử nghiệm **Residual Masking Network** với nhiều kích thước hoặc cấu hình khác nhau (ví dụ, các phiên bản nhẹ hơn với ít tham số hơn hoặc các phiên bản sâu hơn để tăng hiệu suất). Việc điều chỉnh kiến trúc mô hình, chẳng hạn như thay đổi số lớp tích chập, số kênh, hoặc cấu trúc cơ chế chú ý, có thể mang lại cải thiện đáng kể nhưng không được thực hiện do ràng buộc thời gian.
* **Giới hạn về tập dữ liệu**: Tác giả cũng chưa có cơ hội đánh giá mô hình trên các tập dữ liệu lớn hơn hoặc đa dạng hơn ngoài **FER2013** và **VEMO**. Ví dụ, các tập dữ liệu như AffectNet (chứa hơn 1 triệu hình ảnh) hoặc EmotioNet (tập trung vào các biểu cảm phức tạp) có thể cung cấp thêm thông tin về hiệu suất của mô hình trong các điều kiện thực tế hơn, nhưng việc tích hợp chúng đòi hỏi thời gian và nguồn lực vượt ngoài khả năng của luận văn.

Hạn chế này phản ánh thực tế phổ biến trong các nghiên cứu học thuật cấp đại học, nơi thời gian và tài nguyên thường bị giới hạn. Tuy nhiên, nó cũng nhấn mạnh tiềm năng cải tiến của Residual Masking Network nếu được tiếp tục phát triển trong các dự án dài hạn hơn.

## **Số lượng tham số lớn**

**Residual Masking Network**, mô hình chính của luận văn, có **142.9 triệu tham số**, một con số đáng kể so với các mô hình so sánh như **ResNet18** (11.2 triệu tham số) hay **DenseNet121** (6.9 triệu tham số). Số lượng tham số lớn này, mặc dù góp phần vào hiệu suất cao của mô hình, lại đặt ra một số thách thức quan trọng:

* **Yêu cầu tài nguyên tính toán cao**: Với 142.9 triệu tham số, Residual Masking Network đòi hỏi phần cứng mạnh mẽ (như GPU cao cấp hoặc TPU) để huấn luyện và triển khai. Trong bối cảnh nghiên cứu, điều này có thể không phải là vấn đề lớn, nhưng khi xem xét ứng dụng thực tế, đặc biệt tại Việt Nam – nơi nhiều tổ chức và cá nhân vẫn sử dụng thiết bị có hiệu năng trung bình – mô hình trở nên khó triển khai.
* **Thời gian xử lý**: Số lượng tham số lớn dẫn đến thời gian huấn luyện và suy luận (inference) lâu hơn, làm tăng chi phí tính toán và giảm tính khả thi trong các ứng dụng yêu cầu phản hồi thời gian thực, chẳng hạn như giám sát an ninh hoặc giao tiếp người-máy.
* **Khả năng triển khai trên thiết bị yếu**: Các thiết bị nhúng (embedded devices) hoặc thiết bị di động với bộ nhớ và sức mạnh tính toán hạn chế (như điện thoại thông minh hoặc camera IoT) khó có thể chạy Residual Masking Network mà không cần tối ưu hóa đáng kể, chẳng hạn như nén mô hình (model compression) hoặc lượng tử hóa (quantization).

Hạn chế này phản ánh một vấn đề chung trong học sâu: sự đánh đổi giữa hiệu suất và hiệu quả tính toán. Mặc dù Residual Masking Network đạt độ chính xác cao, số lượng tham số lớn có thể giới hạn khả năng ứng dụng của nó trong các kịch bản thực tế, đặc biệt khi tài nguyên phần cứng là một ràng buộc.

## **Chưa kiểm chứng trên các bài toán khác**

Residual Masking Network được thiết kế và thử nghiệm chủ yếu cho bài toán **phân loại cảm xúc khuôn mặt**, với trọng tâm là tập dữ liệu FER2013 và VEMO. Tuy nhiên, tác giả chưa có cơ hội đánh giá hiệu quả của mô hình trên các bài toán thị giác máy tính khác, chẳng hạn như:

* **Phát hiện đối tượng (object detection)**: Xác định vị trí và phân loại các đối tượng trong hình ảnh, chẳng hạn như nhận diện khuôn mặt trước khi phân loại cảm xúc.
* **Phân đoạn hình ảnh (image segmentation)**: Tách biệt các vùng khuôn mặt (như mắt, miệng) để phân tích chi tiết hơn về các đặc trưng cảm xúc.
* **Nhận diện hành động (action recognition)**: Phân tích các chuyển động khuôn mặt theo thời gian từ video, thay vì chỉ xử lý hình ảnh tĩnh.

Việc giới hạn trong bài toán phân loại cảm xúc là hợp lý trong phạm vi một luận văn tốt nghiệp, nhưng nó cũng làm giảm khả năng hiểu đầy đủ về tính linh hoạt của Residual Masking Network. Cơ chế chú ý, vốn là điểm mạnh của mô hình, có tiềm năng được ứng dụng trong nhiều bài toán khác, nhưng việc thiếu thử nghiệm trên các lĩnh vực này khiến giá trị tổng quát của mô hình chưa được khai thác triệt để. Hạn chế này mở ra một hướng nghiên cứu hấp dẫn cho tương lai, nơi Residual Masking Network có thể được điều chỉnh để giải quyết các bài toán thị giác máy tính đa dạng hơn.

## **Hạn chế của tập dữ liệu VEMO**

Tập dữ liệu **VEMO**, mặc dù là một đóng góp quan trọng của luận văn, vẫn tồn tại một số hạn chế liên quan đến quy mô và tính đa dạng:

* **Quy mô chưa đủ lớn**: So với các tập dữ liệu quốc tế như FER2013 (hơn 35,000 hình ảnh) hoặc AffectNet (hơn 1 triệu hình ảnh), VEMO có khả năng nhỏ hơn đáng kể về số lượng mẫu. Một tập dữ liệu nhỏ có thể không đủ để huấn luyện các mô hình học sâu phức tạp như Residual Masking Network một cách tối ưu, dẫn đến hiệu suất thấp hơn (65.949% trên VEMO so với 74.14% trên FER2013).
* **Tính đa dạng hạn chế**: Mặc dù VEMO được thiết kế để đại diện cho người Việt Nam, tập dữ liệu có thể chưa bao quát hết các yếu tố như độ tuổi, giới tính, vùng miền, hoặc bối cảnh văn hóa khác nhau. Ví dụ, cách thể hiện cảm xúc của người miền Bắc có thể khác với người miền Nam, hoặc người trẻ có thể biểu cảm khác so với người lớn tuổi. Những biến thể này đòi hỏi một tập dữ liệu lớn hơn và đa dạng hơn để phản ánh đầy đủ.
* **Chất lượng và gắn nhãn**: Quá trình xây dựng VEMO có thể gặp khó khăn trong việc đảm bảo chất lượng hình ảnh (ví dụ, độ phân giải, điều kiện ánh sáng) và độ chính xác của nhãn cảm xúc. Gắn nhãn thủ công hoặc bán tự động dễ dẫn đến sai sót, đặc biệt với các cảm xúc phức tạp hoặc biểu cảm hỗn hợp.

Những hạn chế này không làm giảm giá trị của VEMO, bởi việc xây dựng một tập dữ liệu mới là một nỗ lực đáng kể, đặc biệt trong bối cảnh Việt Nam thiếu tài nguyên dữ liệu địa phương. Tuy nhiên, chúng chỉ ra rằng VEMO cần được mở rộng và cải thiện về quy mô, tính đa dạng, và chất lượng để trở thành một chuẩn mực đáng tin cậy cho các nghiên cứu nhận diện cảm xúc tại Việt Nam.

## **Khả năng giải thích mô hình**

Mặc dù luận văn đã sử dụng các phương pháp trực quan hóa tiên tiến như **Grad-CAM** và **gộp trung bình đặc trưng** để giải thích cách Residual Masking Network đưa ra dự đoán, khả năng giải thích chi tiết vẫn còn hạn chế, đặc biệt trong các trường hợp mô hình dự đoán sai:

* **Hạn chế của Grad-CAM**: Grad-CAM cung cấp bản đồ nhiệt để chỉ ra các vùng mà mô hình tập trung, nhưng nó không giải thích được tại sao mô hình lại ưu tiên các vùng đó hoặc tại sao một số dự đoán sai xảy ra. Ví dụ, nếu mô hình nhầm lẫn giữa cảm xúc giận dữ và buồn bã, Grad-CAM có thể cho thấy nó tập trung vào vùng mắt, nhưng không làm rõ đặc trưng cụ thể nào dẫn đến sai lầm.
* **Các cảm xúc phức tạp**: Bài toán nhận diện cảm xúc thường gặp khó khăn với các biểu cảm hỗn hợp (mixed emotions) hoặc cảm xúc ít rõ ràng (như sự mỉa mai hoặc bối rối), vốn không được đại diện đầy đủ trong FER2013 hoặc VEMO. Residual Masking Network, dù mạnh mẽ, chưa được tối ưu hóa để xử lý các trường hợp này, và các công cụ trực quan hóa hiện tại không cung cấp đủ thông tin để phân tích sâu hơn.
* **Thiếu phân tích lỗi chi tiết**: Luận văn chưa đi sâu vào việc phân tích các trường hợp dự đoán sai thông qua các kỹ thuật như phân tích lỗi (error analysis) hoặc kiểm tra các mẫu dữ liệu cụ thể. Điều này có thể giúp xác định các yếu tố như nhiễu dữ liệu, nhãn sai, hoặc hạn chế của mô hình trong việc học các đặc trưng phức tạp.

Hạn chế này phản ánh một thách thức chung trong học sâu: việc giải thích các mô hình phức tạp, đặc biệt khi chúng có số lượng tham số lớn như Residual Masking Network. Tuy nhiên, việc sử dụng Grad-CAM đã là một bước tiến đáng kể, và hạn chế này chỉ ra tiềm năng cải thiện thông qua các kỹ thuật giải thích tiên tiến hơn, như SHAP (SHapley Additive exPlanations) hoặc LIME (Local Interpretable Model-agnostic Explanations).

Những hạn chế trên không làm giảm giá trị của luận văn mà thay vào đó cung cấp một bức tranh thực tế về các thách thức trong nghiên cứu nhận diện cảm xúc. Chúng cũng phản ánh sự trung thực của tác giả trong việc đánh giá công trình của mình, một phẩm chất quan trọng trong nghiên cứu khoa học. Cụ thể:

* **Thời gian và tham số**: Các hạn chế về thời gian và số lượng tham số là những vấn đề kỹ thuật có thể được giải quyết trong các dự án dài hạn hơn, với sự hỗ trợ của tài nguyên tính toán tốt hơn hoặc các kỹ thuật tối ưu hóa như cắt tỉa mô hình (pruning) và lượng tử hóa.
* **Ứng dụng đa bài toán**: Việc chưa kiểm chứng trên các bài toán khác không phải là điểm yếu mà là cơ hội để mở rộng Residual Masking Network, vốn có tiềm năng lớn nhờ cơ chế chú ý linh hoạt.
* **Tập dữ liệu VEMO**: Hạn chế của VEMO là một lời nhắc nhở về tầm quan trọng của việc xây dựng dữ liệu địa phương, đồng thời khuyến khích các nỗ lực cộng đồng để mở rộng và cải thiện nó.
* **Giải thích mô hình**: Khả năng giải thích hạn chế là một vấn đề phổ biến trong học sâu, nhưng việc sử dụng Grad-CAM đã đặt nền móng cho các cải tiến tiếp theo, đặc biệt khi các kỹ thuật giải thích AI đang phát triển nhanh chóng.

Các hạn chế này mở ra nhiều hướng nghiên cứu tiềm năng:

* **Tối ưu hóa mô hình**: Phát triển các phiên bản nhẹ hơn của Residual Masking Network để phù hợp với thiết bị yếu, sử dụng các kỹ thuật như nén mô hình hoặc kiến trúc hiệu quả (efficient architectures) như MobileNet.
* **Mở rộng VEMO**: Tăng quy mô và tính đa dạng của VEMO thông qua hợp tác với các tổ chức hoặc sử dụng dữ liệu đám đông (crowdsourcing) để thu thập thêm hình ảnh.
* **Ứng dụng đa dạng**: Thử nghiệm Residual Masking Network trên các bài toán khác như phân đoạn khuôn mặt, nhận diện hành động, hoặc phân tích cảm xúc từ video.
* **Cải thiện giải thích**: Áp dụng các kỹ thuật giải thích tiên tiến hơn và phân tích lỗi chi tiết để hiểu rõ nguyên nhân của các dự đoán sai, đặc biệt với các cảm xúc phức tạp.

# **Kết luận**

Luận văn “Nhận diện cảm xúc mặt người sử dụng mạng học sâu có cơ chế chú ý” của Phạm Quý Luận là một công trình nghiên cứu đáng chú ý, kết hợp giữa lý thuyết học sâu và ứng dụng thực tiễn. Với việc đề xuất Residual Masking Network, xây dựng tập dữ liệu VEMO, và đạt được độ chính xác vượt trội trên cả FER2013 (76.82%) và VEMO (65.949%), nghiên cứu đã đóng góp quan trọng vào lĩnh vực nhận diện cảm xúc tại Việt Nam. Tuy vẫn còn một số hạn chế về thời gian và tài nguyên, luận văn đã đặt nền móng cho các nghiên cứu tiếp theo, từ việc cải tiến mô hình đến ứng dụng thực tế trong các hệ thống thông minh. Công trình này không chỉ thể hiện năng lực nghiên cứu của tác giả mà còn khẳng định vai trò của học sâu trong việc giải quyết các bài toán phức tạp liên quan đến con người.