

ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

TRẦN HOÀNG TUẤN

SINH BIỂU CẢM KHUÔN MẶT DỰA TRÊN
PHÙ HỢP GIỌNG NÓI

Chuyên ngành: Khoa học dữ liệu
Mã số: ...

LUẬN VĂN THẠC SĨ

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI
TRƯỜNG ĐẠI HỌC BÁCH KHOA - ĐHQG - TP.HCM

Cán bộ hướng dẫn khoa học:...

Cán bộ chấm nhận xét 1:

Cán bộ chấm nhận xét 2:

Mục lục

1	Giới thiệu đề tài	2
2	Mục tiêu, giới hạn và đối tượng nghiên cứu	3
2.1	Mục tiêu	3
2.2	Giới hạn	3
2.3	Đối tượng nghiên cứu	3
2.4	Kết quả dự kiến	3
3	Phương pháp nghiên cứu	4
3.1	Phương pháp đánh giá kết quả nghiên cứu	4
3.2	Phương pháp thu thập và phân tích dữ liệu	4
3.3	Các thí nghiệm dự kiến sẽ triển khai	6
4	Kế hoạch triển khai	7
5	Tổng quan các công trình nghiên cứu liên quan	8
5.1	Bài nghiên cứu "Lip Movements Generation at a Glance" ^[1]	8
5.2	Bài nghiên cứu "End-to-End Speech-Driven Facial Animation with Temporal GANs" ^[2]	11
5.3	Bài nghiên cứu "Realistic Speech-Driven Facial Animation with GANs" ^[3]	13
6	Nội dung dự kiến của luận văn	16
7	Kết luận	17

Danh sách hình vẽ

1	Ảnh được cắt từ một video trong tập dữ liệu <i>GRID</i>	5
2	Ảnh được cắt từ các video trong tập dữ liệu <i>CREMA-D</i>	5
3	Ảnh được cắt từ các video trong tập dữ liệu <i>LRW</i>	5
4	Ảnh được cắt từ các video trong tập dữ liệu <i>VoxCeleb</i>	6
5	Kế hoạch thực hiện Luận văn	7
6	Mô hình của bài báo Lip Movements Generation at a Glance	8
7	Phương pháp kết hợp đặc trưng hình ảnh và âm thanh	9
8	GANs Discriminator với 3 loại đặc trưng	9
9	Kết quả đánh giá và so sánh mô hình trong nghiên cứu Lip Movements Generation at a Glance	10
10	Mô hình của nghiên cứu End-to-End Speech-Driven Facial Animation with Temporal GANs	11
11	Kiến trúc bộ Generator	11
12	Kiến trúc bộ Sequence Discriminator	11
13	Kết quả của nghiên cứu End-to-End Speech-Driven Facial Animation with Temporal GANs	13
14	Kiến trúc mạng được cập nhật trong nghiên cứu mới của Vougioukas	14
15	Kiến trúc bộ phân biệt đồng bộ Sync Discriminator	14
16	Miêu tả dữ liệu được đưa vào mạng phân biệt đồng bộ	14
17	Kết quả đo đặc của tác giả	15

1 Giới thiệu đề tài

Bài toán tạo sinh dữ liệu dựa trên những nguồn dữ liệu có tính chất khác nhau đã và đang trở thành xu thế trong những năm trở lại đây. Đây là bài toán có tính cấp bách, mang lại giá trị cao về mặt kiến thức cho ngành trí tuệ nhân tạo nói riêng và giá trị về mặt kinh tế, công nghệ chung cho toàn xã hội xã hội. Bên cạnh đó, việc tạo sinh dữ liệu về con người đã đạt được những tiến bộ vượt bậc, đặc biệt là tạo sinh dữ liệu hình ảnh khuôn mặt người. Trong đề tài này, mục đích nghiên cứu là: cho biết một vài dữ liệu về gương mặt của một người bất kỳ (hình ảnh, video ngắn) và một đoạn tiếng nói bất kỳ, tạo sinh hình ảnh khuôn mặt người đó đang nói đoạn tiếng nói đã cho một cách chân thực.

Ý nghĩa khoa học: Dòng góp cho sự phát triển chung của xu hướng tạo sinh dữ liệu mới dựa trên các tính chất của dữ liệu ban đầu. Việc tìm ra phương pháp giải quyết tốt bài toán sẽ tạo nền tảng để giải quyết những bài toán xa hơn, phức tạp hơn như: tạo sinh nửa người trên, tạo sinh toàn bộ cơ thể người, hay tạo sinh cả một bối cảnh trong phim. Đề tài giúp hiện thực, cải tiến các phương pháp hiện có trong các bài nghiên cứu gần đây, so sánh và cải tiến để cho ra kết quả tạo sinh tốt hơn, đóng góp thêm phương pháp mới cho việc tạo sinh ảnh. Đồng thời, các phương pháp tạo sinh dữ liệu cũng giúp làm giàu dữ liệu để huấn luyện, kiểm thử cho các mô hình học máy, học sâu khác.

Ý nghĩa thực tiễn: Giải quyết thành công vấn đề này đem lại giá trị to lớn về mặt công nghệ, kinh tế và xã hội. Chúng ta có thể tái hiện lại gương mặt người đang nói ở nhiều thứ tiếng khác nhau, tạo sinh khuôn mặt người đại diện trong các hội nghị trực tuyến, tích hợp vào các trò chơi điện tử để làm chúng trở nên chân thực hơn, truyền video trong điều kiện băng thông giới hạn, giả lập trợ lý ảo có hình dáng con người,... Đối với ngành truyền thông, nó có thể tạo ra biên tập viên ảo. Đối với ngành điện ảnh, giải trí, sáng tạo nội dung nó cũng có giá trị ứng dụng khi giúp giảm bớt áp lực lên khâu hóa trang, kỹ xảo.

Kiến trúc mạng Generative Adversarial Network [4] ra đời vào năm 2014 đã đánh dấu một bước chuyển mình mới cho ngành trí tuệ nhân tạo. Kiến trúc này giúp cho việc tạo sinh dữ liệu được thực hiện một cách hiệu quả và chính xác hơn. Dựa trên nền tảng đó, các nghiên cứu về việc tạo sinh ảnh gương mặt người cũng được tiến hành và ngày càng có những bước tiến mới.

Để tạo sinh mặt người đang nói, các công trình nghiên cứu tập trung chủ yếu vào vùng miệng. Bài nghiên cứu vào năm 2018 của Lele Chen [1] đưa ra phương pháp tạo sinh video vùng miệng của người đang nói với đầu vào là ảnh tĩnh của khuôn miệng và một đoạn âm thanh có chứa tiếng nói. Vào năm 2019, Lele Chen [5] và Vougioukas [2] tiếp tục đưa ra phương pháp tạo sinh cả khuôn mặt người dựa vào ảnh tĩnh của khuôn mặt và đoạn âm thanh chứa tiếng nói. Năm 2020, Vougioukas [3] đã cải tiến phương pháp tạo sinh mặt và cập nhật thêm hành động chớp mắt, Lele Chen [6] cũng đưa ra phương pháp mới để tạo sinh mặt hiệu quả hơn, tự nhiên hơn với việc di chuyển của vùng đầu trên khung hình.

Nhìn chung, các nghiên cứu này đã đưa ra các kiến trúc mạng hiệu quả để tạo sinh khuôn mặt cũng như các phương pháp, lập luận và chứng minh tính hiệu quả của các kiến trúc mạng được đề xuất. Mặc dù các thông số của thử nghiệm đưa ra là khá tốt, các nghiên cứu của Vougioukas vẫn chưa thể tạo ra chuyển động của đầu, kết quả tạo sinh của Vougioukas đôi khi không giữ được đặc trưng của ảnh. Nghiên cứu của Lele Chen năm 2020 [6] đã tạo ra chuyển động cho phần đầu dựa trên tiếng nói, nhưng khuôn mặt được tạo sinh vẫn còn có thể bị nhận ra qua các phép thử Turing, và chuyển động của đầu đôi khi vẫn chưa được tự nhiên, mạng cũng có cấu trúc rất phức tạp và đòi hỏi nhiều tài nguyên tính toán để có thể huấn luyện.

2 Mục tiêu, giới hạn và đối tượng nghiên cứu

2.1 Mục tiêu

Mục tiêu của Luận văn Tốt nghiệp là nghiên cứu các đề tài có liên quan bằng cách khảo sát, kiểm định và thử nghiệm các nghiên cứu mới nhất hiện có, qua đó tiến hành các cải tiến, thay đổi và thử nghiệm để đưa ra các kết quả tạo sinh khuôn mặt tốt hơn, tự nhiên hơn, chính xác hơn. Hình ảnh được tạo ra phải sắc nét, ít nhiễu, chân thực và tương đồng về mặt nhận dạng, cấu trúc với hình ảnh người mẫu. Đồng thời, khẩu hình miệng của hình ảnh được tạo ra phải khớp với tiếng nói, phù hợp với cách phát âm từ ngữ. Bên cạnh đó, video được tạo ra phải có tính liền lạc, ổn định, không bị hiện tượng nhảy hình. Mục tiêu được đặt ra nhằm cải thiện các mô hình hiện có, tăng tính ứng dụng của việc tạo sinh mặt người vào thực tiễn cuộc sống.

2.2 Giới hạn

Phạm vi nghiên cứu của Luận văn là tạo sinh ảnh giới hạn trong vùng mặt của người, dữ liệu mẫu được cung cấp ban đầu phải là ảnh rõ ràng của khuôn mặt người, đoạn âm thanh được cung cấp cũng phải là âm thanh rõ ràng của tiếng nói cùng loại với ngôn ngữ được dùng để huấn luyện mạng.

2.3 Đối tượng nghiên cứu

Đối tượng nghiên cứu của Luận văn là các cách tiếp cận, các phương pháp mô hình hóa bài toán, các mạng học máy, học sâu, mạng GANs và các phương pháp tạo sinh dữ liệu từ mạng GANs, các cấu trúc Residual Encoder-Decoder, bên cạnh đó là các phương pháp kết hợp đặc trưng hình ảnh, âm thanh có xem xét đến thứ tự thời gian để tạo sinh hình ảnh mới.

2.4 Kết quả dự kiến

Luận văn sẽ xây dựng được mô hình tính toán mới để có thể tạo sinh hình ảnh mặt người hiệu quả, thỏa mãn được các tiêu chí đã nêu trong phần 2.1. Đồng thời, luận văn cũng sẽ cung cấp được các đánh giá và so sánh khách quan bằng số liệu thực tế.

3 Phương pháp nghiên cứu

3.1 Phương pháp đánh giá kết quả nghiên cứu

Do đặc thù của bài toán là tạo sinh ảnh mặt người, việc lựa chọn phương pháp để đánh giá chất lượng và độ chân thật của hình ảnh được tạo ra là một thách thức lớn. Dựa trên các tiêu chí về mặt hình ảnh được đưa ra ở phần 2.1, đoạn còn lại của phần này sẽ đưa ra các độ đo để đo lường chất lượng hình ảnh của mô hình.

Về chất lượng hình ảnh, hầu hết các nghiên cứu trước đây [1][5][2][3] đều sử dụng *Peak Signal-to-Noise Ratio (PSNR)* và *Structural Similarity Index (SSIM)* như các độ đo để đo lường chất lượng hình ảnh được tạo sinh. Tuy nhiên, PSRN lại không phải là một thông số lý tưởng để đo lường khi nó không quan tâm đến các đặc trưng mặt người trong ảnh. Trong khi đó, SSIM lại là một độ đo tốt hơn khi nó có khả năng đo sự tương đồng về mặt nội dung trong ảnh. SSIM có khả năng so sánh các điểm ảnh và các điểm lân cận của chúng trên ảnh được tạo sinh tương ứng với các điểm trên ảnh thật dựa trên ba tính chất - độ tương phản, độ sáng và cấu trúc. Độ nét cũng là một yếu tố quan trọng nói lên chất lượng ảnh. Độ đo Cummulative Probability Blur Detection được sử dụng trong các nghiên cứu [1][2][3] cũng là một độ đo tốt để đánh giá độ nét của ảnh được tạo sinh.

Để đánh giá việc duy trì nhận dạng người trong các hình ảnh được tạo sinh, *Cosin Similarity (CSIM)* là độ đo được sử dụng để đo lường sự sai lệch về nhận dạng trong ảnh được tạo sinh và ảnh thật. Độ đo này đo lường khoảng cách giữa hai vector đặc trưng của hai ảnh thật và tạo sinh, từ đó cho ra một con số cụ thể về sự sai khác giữa hai ảnh.

Sự phù hợp của khẩu hình miệng với từ ngữ được nói ra cũng là một yếu tố quyết định đối với chất lượng ảnh được tạo sinh. Hiện tại, các nghiên cứu đã đưa ra các độ đo khác nhau để đánh giá sự phù hợp này. Trong [2][3], tác giả sử dụng độ đo *Word Error Rate (WER)*, trong nghiên cứu này, video mặt người sinh ra được đưa vào LipNet - một mô hình mạng học sâu để đọc từ từ khẩu hình miệng, các từ được đọc ra được so sánh với từ gốc và tính ra tỉ lệ sai. Lele Chen [1] đã đưa ra một độ đo khác để đo lường sự phù hợp khẩu hình miệng là *Landmark Distance (LMD)* nhằm đo lường khoảng cách Euclidean giữa các điểm cầu thành khẩu hình miệng trong ảnh thật và ảnh được tạo sinh.

Các độ đo nêu trên là các độ đo đã được kiểm nghiệm tính hiệu quả trong các nghiên cứu gần đây về tạo sinh mặt người dựa vào tiếng nói. Luận văn có thể kế thừa các độ đo này để đánh giá kết quả nghiên cứu, đồng thời có thể dễ dàng so sánh kết quả nghiên cứu của Luận văn với các kết quả nghiên cứu trước đây.

3.2 Phương pháp thu thập và phân tích số liệu

Nhằm mục đích dễ dàng cho việc nghiên cứu, đánh giá và so sánh kết quả với các nghiên cứu trước đây, Luận văn sẽ sử dụng các bộ dữ liệu có sẵn, đã được sử dụng trong các nghiên cứu gần đây có cùng chủ đề. Một số tập dữ liệu sau sẽ được sử dụng trong Luận văn:

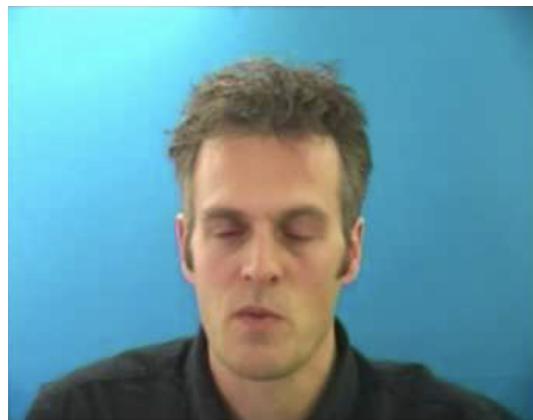
GRID[7]: Được công bố vào năm 2006, GRID là một tập dữ liệu miễn phí nhằm phục vụ mục đích nghiên cứu. Tập dữ liệu chứa video của 34 người (18 nam, 16 nữ), 1000 video mỗi người, mỗi video có độ dài 3 giây và trong video này, người nói chỉ nói một câu duy nhất.

CREMA-D[8]: Được công bố vào năm 2014, CREMA-D là một tập dữ liệu được cung cấp miễn phí cho mục đích nghiên cứu. CREMA-D chứa 7,442 video từ 91 diễn viên, gồm 48 nam và 43 nữ, độ tuổi từ 20 đến 74 từ các chủng tộc người khác nhau. Mỗi diễn viên sẽ nói 12 câu khác nhau, với mỗi câu, họ sẽ thể hiện 6 loại cảm xúc khác nhau khi nói (giận dữ, chán ghét, sợ hãi, vui vẻ, bình thường, và buồn) và 4 cấp độ cảm xúc (thấp, trung bình, cao, và không xác định).

LRW[9]: Là một bộ dữ liệu lớn được thu thập từ kênh truyền hình BBC, LRW chứa đến hơn 1000 giờ video người đang nói, với bộ từ điển hơn 1000 từ vựng, hơn 1 triệu từ đã được nói bởi hơn 1000

người khác nhau.

VoxCeleb[10]: Tập dữ liệu chứa hơn 1000000 từ được nói bởi 1251 người nổi tiếng, những video này được cắt ra từ các video được tải lên YouTube. Tập dữ liệu cũng cân bằng về mặt giới tính với 55% video là nam. Nhân vật trong video đến từ nhiều chủng tộc khác nhau, ngữ điệu khác nhau, ngành nghề và lứa tuổi cũng khác nhau. Video trong tập dữ liệu cũng được thu thập trong nhiều hoàn cảnh khác nhau (trên thảm đỏ, sân vận động, trong phòng thu,...) và tất cả các video đều được ghi bằng các thiết bị cầm tay.



Hình 1: Ảnh được cắt từ một video trong tập dữ liệu *GRID*



Hình 2: Ảnh được cắt từ các video trong tập dữ liệu *CREMA-D*



Hình 3: Ảnh được cắt từ các video trong tập dữ liệu *LRW*



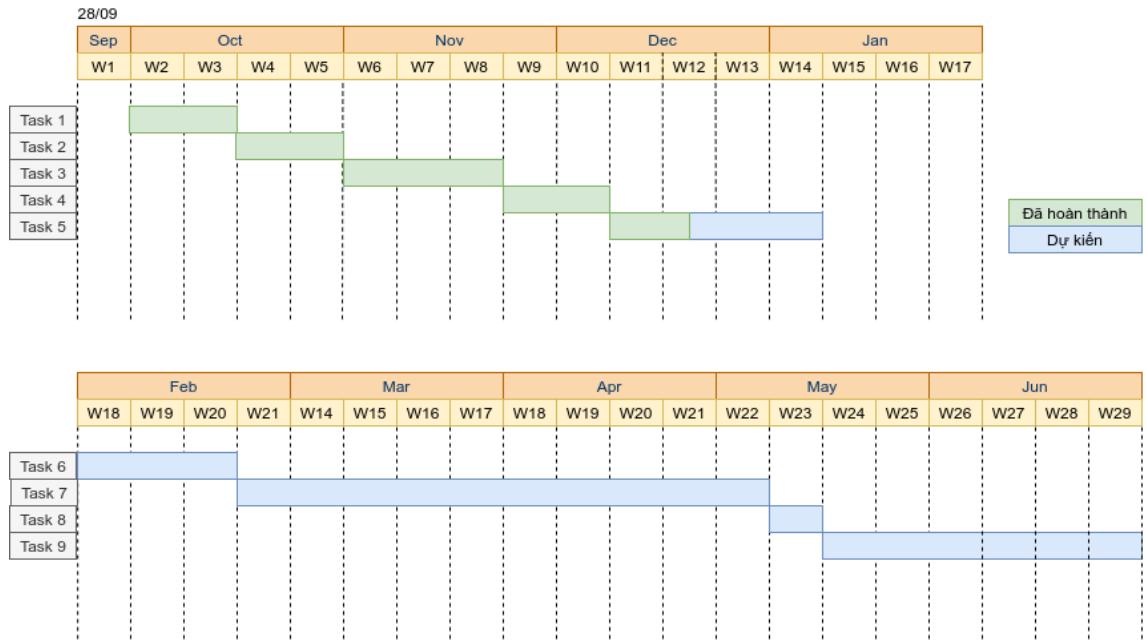
Hình 4: Ảnh được cắt từ các video trong tập dữ liệu *VoxCeleb*

3.3 Các thí nghiệm dự kiến sẽ triển khai

Nhằm mục đích khảo sát, kiểm định và học hỏi, kế thừa các ý tưởng từ các nghiên cứu trước, tác giả sẽ đọc hiểu ý tưởng từ các bài báo từ năm 2018 trở lại đây để nắm rõ lý thuyết và cách vận hành của mô hình được đề ra. Sau đó cài đặt, tiến hành lại các thí nghiệm trên để kiểm chứng và học hỏi thêm kinh nghiệm, đồng thời sẽ cố gắng đề ra những cải tiến cho các mô hình trên nhằm mục đích tăng chất lượng hình ảnh và tính đồng bộ âm thanh cho video được tạo sinh. Từ việc thử nghiệm, cải tiến các mô hình của các nghiên cứu trước, tác giả sẽ thiết kế các mô hình mới và tiến hành thử nghiệm với những cài đặt siêu tham số khác nhau.

4 Kế hoạch triển khai

Kế hoạch của Luận văn được chia thành các Task và được thực hiện theo giản đồ GANTT sau:



Hình 5: Kế hoạch thực hiện Luận văn

Task 1: Tìm hiểu các kiến thức liên quan đến Variational Autoencoder và mạng GANs.

Task 2: Tìm hiểu các kiến thức liên quan, ý tưởng, mô hình của nghiên cứu [1]. Đồng thời cài đặt và chạy thử mô hình.

Task 3: Tìm hiểu các kiến thức liên quan, ý tưởng, mô hình của nghiên cứu [2] và [3]. Đồng thời cài đặt và chạy thử mô hình.

Task 4: Tìm hiểu các kiến thức liên quan, ý tưởng, mô hình của nghiên cứu [5]. Đồng thời cài đặt và chạy thử mô hình.

Task 5: Nghiên cứu các phương pháp tạo sinh ảnh dựa trên đặc trưng ba chiều bao gồm nghiên cứu của Lele chen vào năm 2020 [6]. Viết báo cáo và bảo vệ Đề cương Luận văn.

Task 6: Tìm kiếm và tham khảo thêm các nghiên cứu tạo sinh hình ảnh có quan tâm đến cảm xúc tiếng nói.

Task 7: Thiết kế, hiện thực và kiểm chứng các kiến trúc mạng mới.

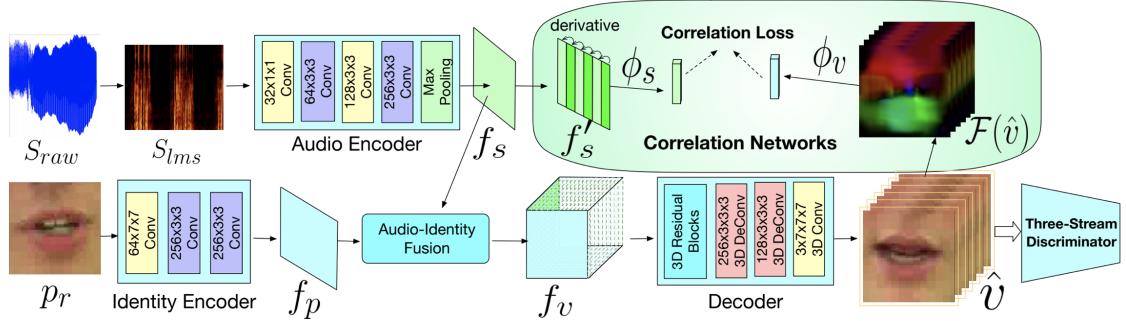
Task 8: Tổng hợp các thử nghiệm và lựa chọn kiến trúc tốt nhất cho Luận văn.

Task 9: Viết báo cáo và bảo vệ Luận văn.

5 Tổng quan các công trình nghiên cứu liên quan

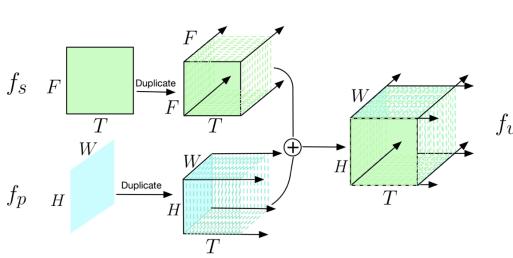
Việc tạo sinh dữ liệu mới trong thời gian gần đây đã phát triển mạnh mẽ với sự ra đời của kiến trúc mạng GANs. Hệ thống mạng GANs bao gồm các mạng học máy, học sâu nhỏ hơn, chia thành hai thành phần là mạng tạo sinh dữ liệu (mạng G) và mạng phân biệt dữ liệu (mạng D). Mạng G đóng vai trò như khối Decoder trong bộ Variational Autoencoder [11], có chức năng học và xấp xỉ được phân phối xác suất của dữ liệu gốc, từ đó tạo sinh ra dữ liệu mới giữ được đặc trưng và tương đồng với dữ liệu gốc. Mạng D có chức năng phân biệt giữa dữ liệu được tạo sinh bởi mạng G và tập dữ liệu huấn luyện. Trong quá trình huấn luyện, dựa trên hàm mất mát của D, các trọng số của cả hai mạng G và D đều được cập nhật trong quá trình lan truyền ngược, từ đó giúp hai mạng này tăng độ chính xác. Với mạng G, qua quá trình huấn luyện, mạng sẽ có khả năng tạo sinh ra được dữ liệu ngày càng chân thực hơn, khó phân biệt hơn. Trong khi đó, mạng D cũng ngày càng có chức năng phân biệt tốt hơn, chuẩn xác hơn. Đến một lúc nào đó, độ chính xác của hai mạng sẽ đạt đến mức cân bằng, lúc này hai mạng đã hội tụ và không thể được cải thiện hơn với kiến trúc mạng và tập dữ liệu huấn luyện hiện tại, nên ta sẽ dừng quá trình huấn luyện tại đây.

5.1 Bài nghiên cứu "Lip Movements Generation at a Glance" [1]

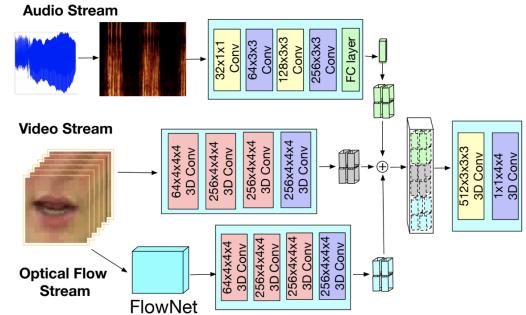


Hình 6: Mô hình của bài báo Lip Movements Generation at a Glance

Việc tạo sinh khung hình miệng khớp với tiếng nói là bước đầu tiên để thực hiện việc tạo sinh khuôn mặt. Nghiên cứu này đã thành công trong việc tạo sinh khung hình miệng từ một ảnh tĩnh chứa hình ảnh khuôn miệng của một người bất kỳ, và một đoạn âm thanh chứa tiếng nói. Bằng phương pháp kết hợp các đặc trưng âm thanh và hình ảnh, nghiên cứu cho ra kết quả tốt và có độ chính xác cao hơn so với các nghiên cứu trước đó. Hình 6 mô tả cấu trúc của mạng tạo sinh ảnh được dùng. Đầu tiên, âm thanh được cắt thành các đoạn nhỏ dài 0.64s, các đoạn này được chuyển thành phổ Log-Mel (S_{raw} thành S_{lms}), sau đó qua một bộ Audio Encoder để trích đặc trưng, ta có đặc trưng âm thanh f_s là một ma trận 2 chiều kích thước $F \times T$. Bên cạnh đó, hình ảnh khuôn miệng cũng được đưa qua một bộ Identity Encoder để tạo thành ma trận 2 chiều f_p kích thước $H \times W$.



Hình 7: Phương pháp kết hợp đặc trưng hình ảnh và âm thanh



Hình 8: GANs Discriminator với 3 loại đặc trưng

Để kết hợp hai đặc trưng hình ảnh, âm thanh với nhau để tạo sinh hình ảnh mới, tác giả bài báo đề xuất phương pháp nhân bản ma trận f_s F lần và nhân bản ma trận f_p T lần thành hai tensor ba chiều, sau đó nối tiếp các kênh của hai tensor này để tạo thành khối tensor ba chiều mới. Để có thể nối tiếp được với nhau, Tác giả đã đặt các thông số $H = W = F$. Hình 7 miêu tả cách kết hợp hai đặc trưng hình ảnh - âm thanh thành khối đặc trưng chung f_v , khối f_v có kích thước $W \times H \times T$. Khối đặc trưng này sau cùng được chuyển đổi thành ảnh đầu ra \hat{v} nhờ vào mạng Decoder. Mạng Decoder này sử dụng kiến trúc 3D Residual và các khối Deconvolution nhằm bảo toàn các đặc điểm của hình ảnh gốc.

Đồng thời, nghiên cứu này cũng chỉ ra rằng đặc tính của khuôn miệng trong video là hình ảnh thường di trước âm thanh, và độ trễ âm thanh - hình ảnh là không đồng nhất trong các video khác nhau. Vì vậy, để tạo sinh một video chân thực, ta phải quan tâm đến độ trễ này. Khối Correlation Network trong hình 6 miêu tả cách tính toán giá trị Corelation Loss. Để tính giá trị này, bộ tính toán cần có một encoder (ϕ_s) để encode sự thay đổi của âm thanh và một encoder (ϕ_v) để encode sự thay đổi của hình ảnh. Ma trận f_s được tính đạo hàm theo trục T thành f'_s , hình ảnh được tạo sinh \hat{v} cũng được đưa qua hàm \mathcal{F} để lấy đặc trưng Optical Flow. Sau đó, cả hai đặc trưng thể hiện sự thay đổi của âm thanh và hình ảnh theo thời gian này được đưa qua các encoder ϕ_s và ϕ_v để tạo ra hai vector có cùng số chiều. Công thức tính của giá trị Correlation Loss chính là hiệu của 1 và giá trị cosin giữa hai vector đặc trưng cuối cùng:

$$\ell_{corr} = 1 - \frac{\phi_s(f'_s) \cdot \phi_v(\mathcal{F}(\hat{v}))}{\|\phi_s(f'_s)\|_2 \cdot \|\phi_v(\mathcal{F}(\hat{v}))\|_2} \quad (1)$$

Cấu trúc mạng GANs cũng đã được sử dụng trong nghiên cứu này nhằm mục đích tạo ra chuyển động mượt mà cho chuỗi hình ảnh trong video và làm cho chất lượng ảnh tạo sinh tốt hơn. Bộ phân biệt (Discriminator) giữa ảnh thật và ảnh tạo sinh (D) được miêu tả trong hình 8. Đặc trưng Log-Mel của âm thanh được encode thành một vector bằng một mạng Convolution - Fully connected, sau đó vector này được nhân bản và ghép nối để có số chiều bằng với tensor của hai đặc trưng còn lại. Hình ảnh được đưa vào mạng được encode bởi các khối 3D Convolution để có được tensor đặc trưng ảnh. Các ảnh này cũng được đưa qua mạng FlowNet để đưa ra đặc trưng Optical Flow, đặc trưng này cũng được encode để tạo ra tensor đặc trưng cho chuyển động trong video. Sau cùng, ba đặc trưng này được ghép nối tiếp theo kênh và được đưa qua các khối 3D Convolution để lấy được xác suất dự đoán ảnh thật hay tạo sinh của mạng. Cặp video - âm thanh đưa vào mạng có thể là video thật và đoạn âm thanh khớp với video đó, hoặc video thật và một đoạn âm thanh khác, hoặc video được tạo sinh và đoạn âm thanh tương ứng tạo ra nó. Hàm mất mát của mạng được định nghĩa như sau:

$$\ell_{dis} = -\log D([s^j, v^j]) - \lambda_p \log(1 - D([s^j, \hat{v}])) - \lambda_u \log(1 - D([s^j, v^k])), k \neq j \quad (2)$$

Để so sánh sự tương đồng về mặt tổng quan giữa hai video (video thật và video được tạo sinh),

tác giả sử dụng một bộ Autoencoder. Bộ Autoencoder này được huấn luyện độc lập với mạng chính và sử dụng cùng bộ dữ liệu với mạng chính. Phần encoder (φ) được giữ lại để encode hình ảnh từ video nhằm mục đích trích xuất đặc trưng của chuỗi hình ảnh. Hàm Perceptual Loss $\ell_{perc}(\hat{v}, v)$ được dùng để tính toán độ sai lệch về mặt tổng quan giữa video được tạo sinh từ tiếng nói và video thật tương ứng với tiếng nói:

$$\ell_{perc}(\hat{v}, v) = \|\varphi(v) - \varphi(\hat{v})\|_2^2 \quad (3)$$

Hàm mất mát của cuối cùng của mạng được định nghĩa:

$$\mathcal{L} = \ell_{corr} + \lambda_1 \ell_{pix} + \lambda_2 \ell_{perc} + \lambda_3 \ell_{gen} \quad (4)$$

Trong đó:

- ℓ_{corr} : Là giá trị mất mát do sự sai lệch giữa hình ảnh và âm thanh đã nêu ở (1).
- ℓ_{pix} : Giá trị mất mát dựa trên sự sai khác ở cấp độ điểm ảnh giữa ảnh được tạo sinh và ảnh trong video thật, $\ell_{pix} = \|v - \hat{v}\|$.
- ℓ_{perc} : Giá trị mất mát đo độ sai khác trên toàn bộ chuỗi hình ảnh (đã nêu ở (3)).
- ℓ_{gen} : Giá trị mất mát của bộ tạo sinh ảnh dựa trên hàm phân biệt D : $\ell_{gen} = -\log D([s^j, \hat{v}^j])$.

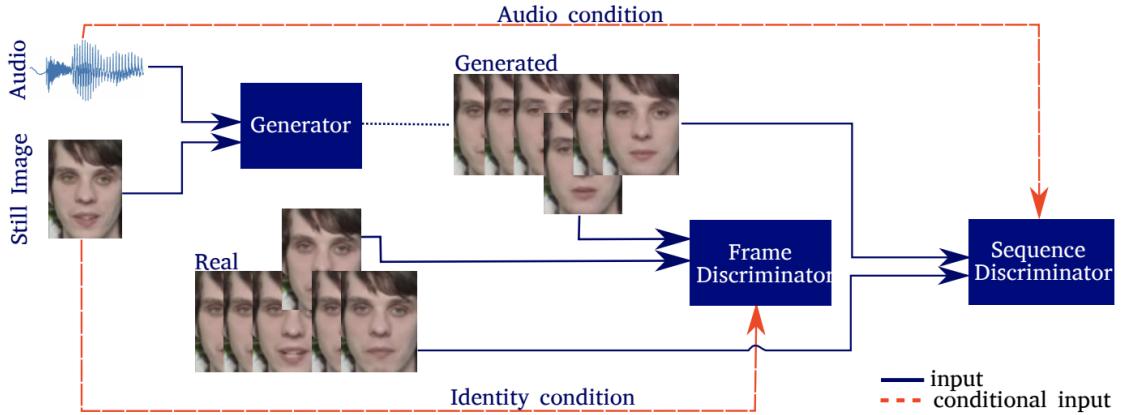
Mô hình được huấn luyện và kiểm thử trên các tập dữ liệu GRID, LDC và LRW. Kết quả kiểm thử cho thấy mô hình này cho kết quả tạo sinh hình ảnh tốt hơn hẳn so với các nghiên cứu trước đó. Các độ đo PSNR, SSIM, LMD và CPBD (đã nêu ở phần 3.1) được sử dụng để kiểm chứng. Sau đây là kết quả được khảo sát bởi tác giả:

Method	GRID				LDC				LRW			
	LMD	SSIM	PSNR	CPBD	LMD	SSIM	PSNR	CPBD	LMD	SSIM	PSNR	CPBD
G. T.	0	N/A	N/A	0.141	0	N/A	N/A	0.211	0	N/A	N/A	0.068
Vondrick[30]	2.38	0.60	28.45	0.129	2.34	0.75	27.96	0.160	3.28	0.34	28.03	0.082
Chung [6]	1.35	0.74	29.36	0.016	2.13	0.50	28.22	0.010	2.25	0.46	28.06	0.083
Full model	1.18	0.73	29.89	0.175	1.82	0.57	28.87	0.172	1.92	0.53	28.65	0.075

Hình 9: Kết quả đánh giá và so sánh mô hình trong nghiên cứu Lip Movements Generation at a Glance

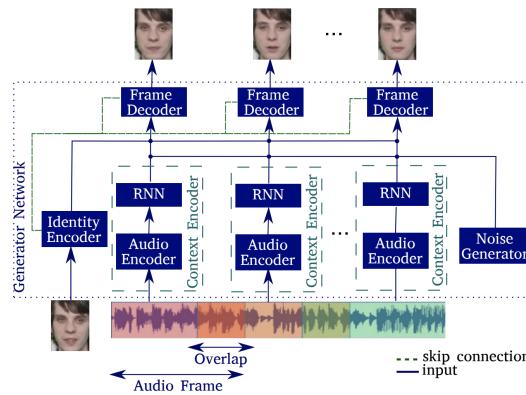
Nghiên cứu này đã đưa ra các phương pháp phù hợp và tiên bộ để trích xuất và kết hợp đặc trưng hình ảnh và âm thanh, đồng thời cũng tận dụng phương pháp GANs để cải thiện chất lượng của ảnh được tạo sinh. Độ hiệu quả của cấu trúc mạng được thể hiện qua kết quả đo lường vượt trội so với các nghiên cứu cùng thời điểm. Tuy nhiên, cấu trúc mạng này có một số yếu điểm. Thứ nhất, mạng chỉ có thể nhận vào hình ảnh tĩnh và một đoạn âm thanh có độ dài xác định (0.64s) và cho ra số khung hình tương ứng với khoảng thời gian đó (16 khung hình). Thứ hai, tác giả vẫn chưa chú ý đến hiện tượng nhảy hình của video được tạo sinh, mạng không có cơ chế để đảm bảo việc chuyển ảnh mượt mà, ít sai khác về độ tương phản, ánh sáng, màu sắc giữa các khung ảnh gần nhau.

5.2 Bài nghiên cứu "End-to-End Speech-Driven Facial Animation with Temporal GANs" [2]

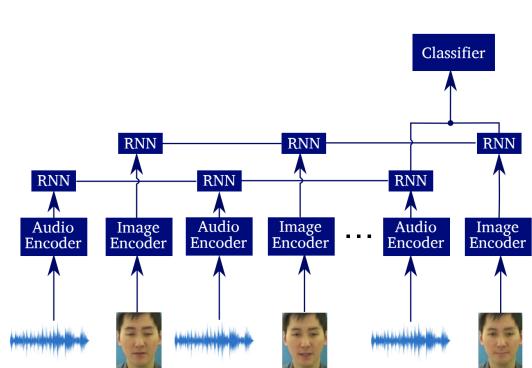


Hình 10: Mô hình của nghiên cứu End-to-End Speech-Driven Facial Animation with Temporal GANs

Nghiên cứu của Vougioukas vào năm 2019 có mục tiêu là tạo ra chuỗi hình ảnh của toàn bộ gương mặt người đang nói với đầu vào là một ảnh tĩnh chứa mặt người bất kỳ và một đoạn tiếng nói bất kỳ. Kiến trúc của mạng được miêu tả trong hình 10, được bao gồm ba phần chính. Bộ tạo sinh ảnh Generator nhận vào một đoạn âm thanh tiếng nói có độ dài bất kỳ và một ảnh tĩnh, sử dụng những dữ liệu này như một gợi ý để tạo sinh chuỗi hình ảnh mới với mặt người đang nói tiếng nói tương ứng. Bộ phân biệt ảnh theo khung hình Frame Discriminator cũng nhận vào ảnh tĩnh nói trên, đồng thời nhận thêm một ảnh khác, ảnh này hoặc được tạo sinh từ Generator, hoặc được trích xuất từ video trong bộ dữ liệu. Frame Discriminator sẽ được huấn luyện để phân biệt đâu là ảnh được tạo sinh và đâu là ảnh thật được trích xuất từ video huấn luyện. Bộ phân biệt chuỗi ảnh Sequence Discriminator nhận vào đoạn tiếng nói và một chuỗi hình ảnh trong video. Chuỗi hình ảnh này có thể là hình ảnh được tạo sinh hoặc hình ảnh gốc từ video tương ứng. Bộ Sequence Discriminator sẽ học cách phân biệt hai loại video này trong quá trình huấn luyện. Theo như cơ chế GANs, trong quá trình huấn luyện, hàm mất mát của hai bộ Discriminator sẽ tạo ra lan truyền ngược và cập nhật trọng số của chính nó và của cả Generator. Từ đó, cả ba mạng này đều sẽ dần trở nên chính xác hơn.



Hình 11: Kiến trúc bộ Generator



Hình 12: Kiến trúc bộ Sequence Discriminator

Trong nghiên cứu này, video huấn luyện được tách riêng thành âm thanh và hình ảnh. Mỗi khung hình kích thước 96×128 được cắt ra từ video tương ứng với đoạn âm thanh dài 0.16s, đoạn âm thanh này tạo thành một vector 8000 điểm. Như vậy, dữ liệu được tiền xử lý bằng cách cắt nhỏ video thành khung ảnh và đoạn âm thanh tương ứng với ảnh đó (âm thanh có chồng lấn giữa các ảnh).

Kiến trúc của bộ tạo sinh chuỗi ảnh Generator được miêu tả trong hình 11. Identity Encoder là bộ encoder ảnh có tính năng trích xuất đặc trưng của ảnh tĩnh được đưa vào mạng. Bộ encoder này gồm 6 lớp 2D Convolution, mỗi lớp kết hợp với Batchnorm và ReLU ở phía sau. Mạng này giúp trích xuất ảnh đầu vào thành vector 50 chiều (z_{id}). Context Encoder là tổng hợp của hai bộ bao gồm Audio Encoder và một bộ RNN hai lớp. Audio Encoder sẽ trích xuất đặc trưng từ đoạn âm thanh 8000 điểm (vector 8000 chiều) để tạo ra vector 256 chiều. Đặc trưng này được đưa vào bộ RNN để bổ sung ngữ nghĩa về mặt thời gian. Ngõ ra của bộ Context Encoder là trạng thái ẩn (hidden state) của bộ RNN có số chiều bằng 256 (z_c). Generator còn có một bộ Noise Generator, thực chất là một mạng GRU có chức năng tạo ra vector nhiễu Gauss 10 chiều (z_n). z_{id} , z_c và z_n được ráp nối với nhau theo kenh tương ứng trước khi đưa vào bộ tạo sinh ảnh. Trong khi z_{id} có chức năng giúp Frame Decoder tái tạo chính xác gương mặt của người nói, z_c sẽ mang thông tin về mặt âm thanh, thời gian và hoàn cảnh, giúp tạo ra gợi ý cho mạng để tạo sinh được hình ảnh tương ứng với âm thanh. Đồng thời, z_n tạo ra tính ngẫu nhiên cho mạng, khi đưa cùng một đầu vào, thì sẽ không khi nào mạng cho ra kết quả giống nhau ở hai lần thử. Đồng thời, tính ngẫu nhiên này lại có tính chất phụ thuộc thời gian (do được tạo ra bởi mạng GRU) đem lại cho hình ảnh được tạo sinh các biểu cảm nhỏ như nháy mắt và các chuyển động nhỏ trên mặt một cách liền lạc.

Ở đầu ra, bộ sinh ảnh Frame Decoder được sử dụng để tạo sinh chuỗi ảnh theo thời gian. Vector đặc trưng ẩn có 316 chiều, từ vector đặc trưng này, Frame Decoder sẽ tạo ra hình ảnh có kích thước bằng với hình mẫu ban đầu (96×128). Nhằm bảo toàn nhận dạng của người trong ảnh mẫu, Frame Decoder được thiết kế theo kiến trúc U-Net gồm 6 lớp Convolution tương ứng với Identity Encoder. Các lớp Convolution này nhận thêm các đặc trưng ẩn từ lớp tương ứng của Identity Encoder để hạn chế việc đánh mất nhận dạng của mặt người mẫu do bị ảnh hưởng bởi độ sâu của mạng. Các đặc trưng ẩn đi qua các lớp Deconvolution và cuối cùng tạo ra ảnh tương ứng với âm thanh.

Bộ phân biệt khung ảnh Frame Discriminator trong hình 10 là một mạng Convolution 6 lớp, đầu ra của mạng là xác suất ảnh này được cho là ảnh được tạo sinh. Frame Discriminator giúp cho ảnh tạo sinh từ Generator chân thực hơn, khó phân biệt với ảnh từ video gốc hơn. Bộ phân biệt chuỗi ảnh Sequence Discriminator được miêu tả trong hình 12 có cấu trúc trích xuất đặc trưng tương tự như bộ Generator. Sự khác biệt đến từ bộ trích xuất đặc trưng chuỗi ảnh. Chuỗi hình ảnh được đưa qua bộ Image Encoder để trích xuất đặc trưng ảnh và thu nhỏ số chiều dữ liệu. Các ảnh sau khi qua bộ Image Encoder sẽ được đưa vào mạng RNN hai lớp để cập nhật trạng thái ẩn của mạng RNN. Khi kết thúc chuỗi hình ảnh và âm thanh tương ứng với nó, trạng thái ẩn của hai mạng RNN cho âm thanh và RNN cho hình ảnh được ghép nối tiếp vào nhau theo kenh. Lúc này, đặc trưng âm thanh giúp làm điều kiện để phân biệt chuỗi ảnh tốt hơn. Một bộ Classifier được sử dụng để tính toán xác suất chuỗi ảnh được đưa vào có phải chuỗi được tạo sinh hay không. Bộ Sequence Discriminator giúp cho video tạo ra có sự chân thật trong các chuyển động của khuôn mặt, cũng như sự chân thật trong sự chuyển tiếp giữa các khung hình, nhờ đó tránh được hiện tượng nhảy hình bất thường.

Trong quá trình huấn luyện, các ảnh từ video được cho vào Frame Discriminator (D_{img}) bằng cách lấy mẫu với xác suất đều qua hàm $S(x)$ trên chuỗi ảnh x . Sequence Discriminator (D_{seq}) sẽ phân biệt cả chuỗi ảnh x và âm thanh a . Hàm mất mát của GANs được biểu diễn như sau:

$$\begin{aligned} \mathcal{L}_{adv}(D_{img}, D_{seq}, G) = & E_{x \sim P_d} [\log D_{img}(S(x), x_1)] + E_{z \sim P_z} [\log(1 - D_{img}(S(G(z)), x_1))] + \\ & E_{x \sim P_d} [\log D_{seq}(x, a)] + E_{z \sim P_z} [\log(1 - D_{seq}(G(z), a))] \end{aligned} \quad (5)$$

Hàm mất mát L1 cũng được dùng trên một nửa dưới của ảnh để đảm bảo ảnh được tạo sinh có hình ảnh thể hiện chân thật khuôn miệng và khẩu hình miệng phù hợp với lời nói. Hàm mất mát L1

được biểu diễn như sau:

$$\mathcal{L}_{L1} = \sum_{p \in [0, W] \times [\frac{H}{2}, H]} |F_p - G_p| \quad (6)$$

Như vậy, mục tiêu của cả hệ thống là giảm thiểu hàm mất mát chung bằng cách điều chỉnh các trọng số của bộ tạo sinh ảnh Generator (G) và các bộ phân biệt ảnh Discriminator (D). Hàm mục tiêu của mạng được biểu diễn như sau:

$$\arg \min_G \max_D \mathcal{L}_{adv} + \lambda \mathcal{L}_{L1} \quad (7)$$

Sau đây là bảng so sánh của tác giả với các thông số PSNR, SSIM, CPBD, WER và một số độ đo khác. Bài nghiên cứu cũng so sánh kết quả của họ với một nghiên cứu trước đó (Baseline):

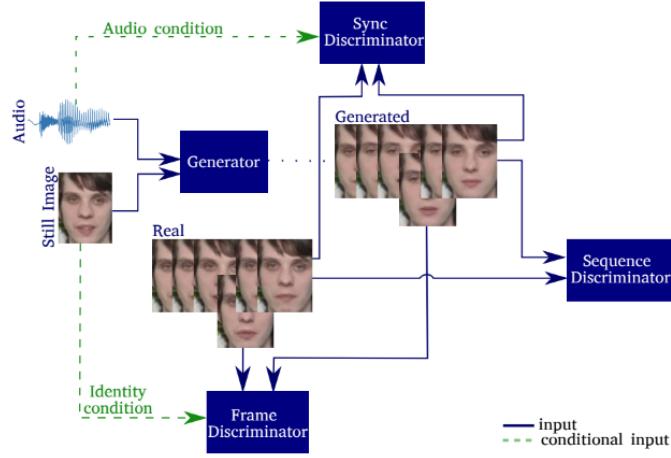
	Method	PSNR	SSIM	FDBM	CPBD	ACD	User	WER
GRID	Proposed Model	27.98	0.844	0.114	0.277	$1.02 \cdot 10^{-4}$	79.77%	25.4%
	Baseline	27.39	0.831	0.113	0.280	$1.07 \cdot 10^{-4}$	20.22%	37.2%
TCD	Proposed Model	23.54	0.697	0.102	0.253	$2.06 \cdot 10^{-4}$	77.03%	N/A
	Baseline	23.01	0.654	0.097	0.252	$2.29 \cdot 10^{-4}$	22.97%	N/A

Hình 13: Kết quả của nghiên cứu End-to-End Speech-Driven Facial Animation with Temporal GANs

Mô hình trong bài nghiên cứu đem lại tín hiệu khả quan cho việc tạo sinh ảnh khuôn mặt dựa trên tiếng nói. Phương pháp tạo sinh và các bộ phân biệt ảnh áp dụng phương pháp CGANs một cách hiệu quả nhằm mục đích tạo sự chân thật cho chuỗi ảnh. Một số kết quả được công bố bởi tác giả bài nghiên cứu cho thấy ảnh được tạo sinh có chất lượng tốt, không bị hiện tượng nhảy hình, độ ổn định khung hình tốt, có thể đánh lừa người xem qua phép thử Turing. Theo đó, có tới 79.77% chuỗi hình ảnh bị đánh nhận sai (được tạo sinh hay video thật) trên tập dữ liệu GRID và 77.03% trên tập dữ liệu TCD. Tuy nhiên, một số hạn chế vẫn chưa được giải quyết. Thứ nhất, ngoài vùng miệng, bộ tạo sinh hình ảnh và các bộ phân biệt vẫn chưa chú trọng đến các phần khác trong khuôn mặt nhất là phần nửa trên. Điều này làm cho hình ảnh được tạo sinh thiếu tự nhiên so với video thực tế. Thứ hai, khuôn mặt vẫn chưa thể hiện được cảm xúc tương ứng với tiếng nói. Việc này cũng góp phần làm cho video được tạo sinh dễ bị nhận biết bởi người xem tinh ý. Thứ ba, chuyển động của đầu cũng không được thể hiện trong video làm cho video trở nên cứng nhắc và giả tạo nếu chiếu trong thời gian dài. Thứ tư, sự đồng bộ của tiếng nói và hình ảnh chưa được quan tâm và chưa có cơ chế đảm bảo hình ảnh được tạo sinh sẽ được căn giờ chuẩn xác với âm thanh.

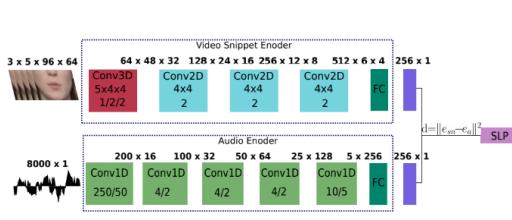
5.3 Bài nghiên cứu "Realistic Speech-Driven Facial Animation with GANs"^[3]

Đây là bài nghiên cứu có cùng tác giả với bài nghiên cứu trong phần 5.2. Với cùng mục tiêu và phương pháp tiếp cận tương đồng với nghiên cứu được công bố năm 2019, Vougioukas đã có một số cập nhật, bổ sung và đánh giá cho mô hình được xây dựng. Trong nghiên cứu này, Vougioukas đã thêm vào mạng trước đó một bộ phân biệt mới, bộ phân biệt này giúp đảm bảo sự đồng bộ giữa hình ảnh được tạo sinh và tiếng nói tương ứng. Kiến trúc được cập nhật mới thể hiện ở hình 14.

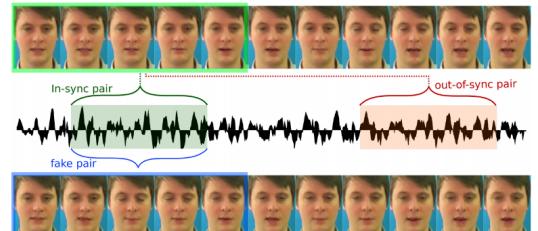


Hình 14: Kiến trúc mạng được cập nhật trong nghiên cứu mới của Vougioukas

Bộ phân biệt tính đồng bộ Sync Discriminator được thể hiện trong hình 15. Bộ phân biệt này nhận vào một chuỗi hình ảnh nửa dưới (phần ảnh chứa vùng miệng) của mặt người đang nói. Chuỗi hình ảnh này có thể được tạo sinh bởi mạng sinh ảnh Generator hoặc là chuỗi ảnh thật được trích xuất từ video. Đồng thời, mạng cũng nhận vào đoạn tiếng nói tương ứng với chuỗi hình ảnh trên. Dữ liệu đưa vào mạng được thể hiện trong hình 16. Chuỗi hình ảnh được đưa vào gồm 5 ảnh gồm 96×64 điểm ảnh tạo thành một tensor ba chiều. Tensor này được biến đổi thành ma trận hai chiều nhờ một lớp 3D Convolution. Sau đó, ma trận này tiếp tục được trích xuất đặc trưng nhờ ba lớp 2D Convolution và ở cuối là một lớp tuyến tính. Đặc trưng ảnh sau cùng được trích xuất là một vector 256 chiều. Đối với tiếng nói, quy trình trích xuất đặc trưng cũng được áp dụng trên vector âm thanh 8000 chiều. Bộ trích xuất đặc trưng âm thanh được sử dụng bao gồm 5 lớp 1D Convolution và cuối cùng là một lớp tuyến tính. Đặc trưng âm thanh cũng được trích xuất thành một vector 256 chiều. Khoảng cách Euclidean giữa hai vector được tính toán theo công thức $d = \|e_{sm} - e_a\|$, với e_{sm} và e_a lần lượt là vector đặc trưng chuỗi hình ảnh và vector đặc trưng tiếng nói. Khoảng cách d sau đó được đưa vào một lớp Perceptron để đưa ra sự đo đạc về độ phù hợp giữa chuỗi hình ảnh vùng miệng và âm thanh được đưa vào dưới dạng xác suất.



Hình 15: Kiến trúc bộ phân biệt đồng bộ Sync Discriminator



Hình 16: Miêu tả dữ liệu được đưa vào mạng phân biệt đồng bộ

Để có khả năng phân biệt tốt nhất, cặp hình ảnh - tiếng nói được đưa vào mạng được chọn lựa theo ba hoàn cảnh (xem hình 16). Cặp đồng bộ đúng (in-sync pair), là cặp hình ảnh - tiếng nói tương ứng với nhau trong video. Cặp không đồng bộ (out-of-sync pair) gồm hình ảnh được trích xuất từ video nhưng không phù hợp với tiếng nói. Cuối cùng là cặp giả mạo (fake pair) gồm hình ảnh được tạo sinh vào tiếng nói tương ứng tạo sinh ra nó. Nhờ ba cặp này, mạng sẽ học được cách phân biệt sự sai lệch hình ảnh - tiếng nói và đặc biệt là sự đồng bộ của khẩu hình miệng với tiếng nói được cải thiện đáng kể.

Hàm mất mát phân biệt \mathcal{L}_{adv} của mạng cũng được cập nhật thêm hàm mất mát của bộ phân biệt

đồng bộ. Qua đó:

$$\mathcal{L}_{adv} = \lambda_{img}\mathcal{L}_{adv}^{img} + \lambda_{adv}\mathcal{L}_{adv}^{seq} + \lambda_{seq}\mathcal{L}_{adv}^{seq} \quad (8)$$

Với:

$$\mathcal{L}_{adv}^{img} = \text{E}_{x \sim P_d}[\log D_{img}(S(x), x_1)] + \text{E}_{z \sim P_z}[\log(1 - D_{img}(S(G(z)), x_1))] \quad (9a)$$

$$\begin{aligned} \mathcal{L}_{adv}^{seq} = & \text{E}_{x \sim P_d}[\log D_{sync}(p_{in})] + \frac{1}{2}\text{E}_{x \sim P_d}[\log(1 - D_{sync}(p_{out}))] + \\ & \frac{1}{2}\text{E}_{z \sim P_z}[\log(1 - D_{sync}(S_{snip}(p_f)))] \end{aligned} \quad (9b)$$

$$\mathcal{L}_{adv}^{seq} = \text{E}_{x \sim P_d}[\log D_{seq}(x, a)] + \text{E}_{z \sim P_z}[\log(1 - D_{seq}(G(z), a))] \quad (9c)$$

Bài nghiên cứu này đã mang lại cải tiến đáng kể cho nghiên cứu trước đó của tác giả. Đặc biệt là cải tiến về mặt đồng bộ về hình ảnh và âm thanh. Điều này giúp cho khẩu hình miệng trở nên chân thật và khớp với hình ảnh được thể hiện qua việc giảm đáng kể độ đo WER (đo độ sai sót của mô hình LipNet - mô hình đọc hình ảnh để đoán từ đang được nói). Các cử động nhỏ trên gương mặt cũng được tái hiện một cách tự nhiên hơn. Bảng sau là kết quả được khảo sát và đo đạc bởi tác giả Vougioukas trên bốn tập dữ liệu GRID, TCD, CREMA và LRW được xử lý bởi bốn mô hình khác nhau:

	Method	PSNR	SSIM	CPBD	ACD	WER	AV Off.	AV Conf.	blinks/sec	blink dur. (sec)
GRID	Proposed Model	27.100	0.818	0.268	$1.47 \cdot 10^{-4}$	23.1%	1	7.4	0.45	0.36
	Baseline	27.023	0.811	0.249	1.42 \cdot 10^{-4}	36.4%	2	6.5	0.04	0.29
	Speech2Vid	22.662	0.720	0.255	$1.48 \cdot 10^{-4}$	58.2%	1	5.3	0.00	0.00
TCD	Proposed Model	24.243	0.730	0.308	1.76 \cdot 10^{-4}	N/A	1	5.5	0.19	0.33
	Baseline	24.187	0.711	0.231	$1.77 \cdot 10^{-4}$	N/A	8	1.4	0.08	0.13
	Speech2Vid	20.305	0.658	0.211	$1.81 \cdot 10^{-4}$	N/A	1	4.6	0.00	0.00
CREMA	Proposed Model	23.565	0.700	0.216	1.40 \cdot 10^{-4}	N/A	2	5.5	0.25	0.26
	Baseline	22.933	0.685	0.212	$1.65 \cdot 10^{-4}$	N/A	2	5.2	0.11	0.13
	Speech2Vid	22.190	0.700	0.217	$1.73 \cdot 10^{-4}$	N/A	1	4.7	0.00	0.00
LRW	Proposed Model	23.077	0.757	0.260	$1.53 \cdot 10^{-4}$	N/A	1	7.4	0.52	0.28
	Baseline	22.884	0.746	0.218	1.02 \cdot 10^{-4}	N/A	2	6.0	0.42	0.13
	Speech2Vid	22.302	0.709	0.199	$2.61 \cdot 10^{-4}$	N/A	2	6.2	0.00	0.00
	ATVGNet	20.107	0.743	0.189	$2.14 \cdot 10^{-4}$	N/A	2	7.0	0.00	0.00

Hình 17: Kết quả đo đạc của tác giả

Qua bảng này, ta thấy độ đo WER là rất thấp so với các mô hình khác nhờ vào sự phù hợp của vùng miệng so với ngữ điệu và nội dung tiếng nói. Các độ đo cơ bản khác như PSNR, SSIM và CPBD cũng được cải thiện so với các mô hình khác. Độ lệch về hình ảnh - âm thanh (AV Off.) cũng rất nhỏ (1). Điểm tự tin về sự đồng bộ giữa hình ảnh - âm thanh (AV Confident) được cải thiện rõ rệt và là bước tiến lớn so với các mô hình khác. Chuyển động của mắt cũng được đo đạc về số lần nháy mắt mỗi giây và thời lượng của một cái nháy mắt. Chuyển động của mắt trong video cũng là rất tự nhiên và đúng với chuyển động nháy mắt của người về tần số và thời lượng. Ngoài các ưu điểm kể trên, một số yếu điểm từ nghiên cứu cũ được nêu ở phần 5.2 vẫn chưa được giải quyết. Đó là các yếu điểm về mặt biểu cảm trên gương mặt và chưa thể hiện được chuyển động của đầu. Ngoài ra, nghiên cứu cũng chưa chú trọng việc tái hiện lại môi trường xung quanh trong video. Qua kiểm nghiệm thực tế, mô hình còn cho thấy một yếu điểm khác khi không giữ được đặc điểm gương mặt của người mẫu nếu hình ảnh người này không tồn tại trong tập dữ liệu huấn luyện. Tác giả đã chạy thử mô hình với hình ảnh người châu Á (mô hình được huấn luyện với bộ dữ liệu mặt người phương Tây), mô hình cho ra mặt người đang nói với đặc trưng gương mặt không còn giống với người châu Á và không giữ được các đường nét đặc trưng của gương mặt người mẫu.

6 Nội dung dự kiến của luận văn

Luận văn tốt nghiệp sẽ được chia thành các phần như sau:

Lời cam đoan của tác giả: Cam đoan các công việc, thử nghiệm và kết quả được đưa ra trong luận văn là trung thực, khách quan.

Tóm tắt luận văn: Trình bày ngắn gọn về cấu trúc của Luận văn, giới thiệu những điểm nhấn của Luận văn, kết quả, và các từ khóa đi kèm.

Mở đầu: Nêu lý do chọn đề tài, mục đích, đối tượng và phạm vi nghiên cứu, ý nghĩa khoa học và ý nghĩa thực tiễn của đề tài.

Tổng quan tình hình nghiên cứu, mục tiêu và nhiệm vụ nghiên cứu: Sơ lược, phân tích, đánh giá các công trình nghiên cứu nổi tiếng có liên quan đến đề tài. Nêu những vấn đề bức thiết cần phải giải quyết, chỉ ra những thiếu sót mà những nghiên cứu trước đây chưa giải quyết được.

Cơ sở lý thuyết: Trình bày cơ sở lý thuyết, các lập luận, căn cứ khoa học được sử dụng trong Luận văn.

Phương pháp nghiên cứu: Trình bày chi tiết về ý tưởng, các mô hình toán, các chứng minh nếu có. Đồng thời trình bày các bước thực hiện và khảo sát, kiểm nghiệm kết quả nghiên cứu. Mô tả kết quả nghiên cứu khi thử nghiệm với nhiều tập dữ liệu và những độ khó khác nhau.

Kết quả nghiên cứu: Mô tả ngắn gọn các kết quả nghiên cứu, thực nghiệm. Bàn luận về điểm mạnh, điểm yếu của mô hình được xây dựng trong luận văn. So sánh kết quả thu được trong quá trình nghiên cứu, thực nghiệm của đề tài và đối chiếu với kết quả nghiên cứu, thực nghiệm của các tác giả khác một cách khách quan. Nêu lên điểm nổi bật, khác biệt của luận văn đối với các nghiên cứu khác.

Kết luận và hướng nghiên cứu mở rộng đề tài: Mô tả, bình luận ngắn gọn và đưa ra kết luận về kết quả nghiên cứu của luận văn và cách thức áp dụng thực tiễn. Đề ra các hướng nghiên cứu mở rộng cho Luận văn.

Danh mục tài liệu tham khảo: Trích dẫn các tài liệu được sử dụng trong Luận văn.

7 Kết luận

Qua đề cương Luận văn, tác giả đã tìm hiểu thêm nhiều kiến thức và học hỏi được các ý tưởng, phương pháp mà các nghiên cứu khác đã áp dụng. Qua đó đề ra một số ý tưởng nghiên cứu riêng cho Luận văn. Luận văn tốt nghiệp sẽ được thực hiện nhằm mục đích giải quyết các vấn đề còn tồn tại của việc tạo sinh hình ảnh gương mặt người đang nói, cố gắng cải thiện chất lượng hình ảnh và thêm vào đó các chuyển động khác như chuyển động của đầu, tóc hay các cơ trên mặt theo cảm xúc người nói. Luận văn cũng sẽ đề ra cách để cải thiện vùng hình ảnh ngoài gương mặt để video được tạo ra chân thực nhất có thể. Để làm được điều này, Luận văn sẽ xem xét và phân tích gương mặt trong không gian ba chiều thay vì hai chiều như trước đây, vì không gian ba chiều cho phép giả lập chuyển động của đầu. Đồng thời cũng sẽ thiết kế một kiến trúc mới có chức năng học và tạo sinh các thông số để sinh ảnh cho một người nhất định.

Tài liệu

- [1] Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 538–553. Springer, 2018.
- [2] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 37–40. Computer Vision Foundation / IEEE, 2019.
- [3] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *Int. J. Comput. Vis.*, 128(5):1398–1413, 2020.
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.
- [5] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7832–7841. Computer Vision Foundation / IEEE, 2019.
- [6] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, volume 12354 of *Lecture Notes in Computer Science*, pages 35–51. Springer, 2020.
- [7] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120 5 Pt 1:2421–4, 2006.
- [8] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.
- [9] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016.
- [10] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.
- [11] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019.