**TÌM LẠI NỤ CƯỜI XƯA**

HỒ VÕ HOÀNG DUY 1, VÕ QUỐC HUY 1, ĐOÀN ÁNH DƯƠNG 1,NGUYỄN XUÂN GIANG 1, NGUYỄN QUANG BẢO 1, DƯƠNG QUANG HUY 1, ĐỖ THU ĐÔNG 1, BÙI THANH HÙNG 1

*1 Khoa công nghệ Thông tin, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh*

*h.hoangduy2002@gmail.com,* *huyvo8500@gmail.com, duongdoan1710@gmail.com, nxgiang2202@gmail.com,nguyenquangbao08@gmail.com, dhuy9954@gmail.com, buithanhhung@iuh.edu.vn.*

**Tóm tắt:** Tạo video từ hình ảnh khuôn mặt có ý nghĩa quan trọng trong việc tạo ra các video tự nhiên từ một bức ảnh, được ứng dụng rộng rãi trong nhiều lĩnh vực như làm phim ảnh hay truyền thông xã hội. Các phương pháp trước đó, còn tồn tại một số hạn chế như tạo ra các video có chất lượng tốt nhưng thiếu tự nhiên trong việc tái tạo lại cử chỉ của nhân vật, một số nghiên cứu chỉ tập trung vào việc tạo các video chất lượng cao mà làm mất đi sự đa dạng về nội dung và phong cách của bức ảnh. Trong nghiên cứu này chúng tôi đề xuất phương pháp tạo ra một video ngắn có chuyển động tự nhiên của môi, mắt và các phần có liên quan trên khuôn mặt bằng cách sử dụng kỹ thuật học sâu, mạng nơ ron tích chập CNN, phép chuyển dổi Hidden Affine kết hợp với Conditional Generative Adversarial Network (cGAN) và một số phương pháp về xử lý ảnh và thị giác máy tính. Chúng tôi sẽ đánh giá trên các bộ dữ liệu CK-Mixed và so sánh với các phương pháp nghiên cứu khác. Dựa trên những kết quả đạt được chúng tôi sẽ tiến hành xây dựng ứng dụng có thể tạo một video có chuyển động khuôn mặt từ một bức ảnh đầu vào và đưa vào thử nghiệm ứng dụng trong thực tiễn.

**FINDING A SMILE OF THE PAST**

**Abstract:** Creating a video from facial images plays a significant role in generating natural-looking videos from a single image. This technique is widely used in various fields such as filmmaking and social media. Previous methods have had limitations, such as creating high-quality videos lacking naturalness in reproducing character movements. Some studies have focused solely on producing high-quality videos, resulting in a loss of diversity in the content and style of the image. In this study, we propose a method to create a short video with natural facial movements of the lips, eyes, and related facial parts using deep learning techniques, convolutional neural networks (CNN), Hiden Affine transformation combined with Conditional Generative Adversarial Network (cGAN), image processing technique and computer vision methods. We will evaluate this method on CK-Mixed datasets and compare it with other research methods. Based on the results, we will proceed to develop an application that can create a facial motion video from a single input image and test its practicality.

**Keywords:** Smiling face video, Deep learning, CNN, Hiden Affine transformation , cGAN, Computer vision.

**1 GIỚI THIỆU**

Trong những năm gần đây, lĩnh vực thị giác máy tính đã chứng kiến những tiến bộ đáng kể, cho phép máy tính nhận biết và hiểu thông tin hình ảnh với độ chính xác ngày càng cao. Một ứng dụng hấp dẫn của công nghệ thị giác máy tính là khả năng chuyển đổi hình ảnh tĩnh thành video động.

Chuyển đổi hình ảnh tĩnh thành video động mang lại những tiềm năng ứng dụng rộng rãi trong lĩnh vực tạo nội dung đa phương tiện, thực tế ảo, chỉnh sửa video và các công nghệ hỗ trợ, làm giàu trải nghiệm người xem và mở ra những khả năng sáng tạo mới. Đồng thời, cũng góp phần mở ra những cánh cửa mới trong việc khai thác tiềm năng của công nghệ thị giác máy tính để tạo ra nội dung sáng tạo và thu hút.

Trong nghiên cứu này, chúng tôi tập trung vào việc tạo ra một video từ một bức ảnh duy nhất, bao gồm các chuyển động của môi, mắt và các phần có liên quan trên khuôn mặt. Chúng tôi đề xuất một phương pháp sáng tạo và hiệu quả để tổng hợp video động từ một bức ảnh tĩnh, tích hợp chuyển động môi, mắt và các phần có liên quan trên khuôn mặt vào video tổng hợp để tạo ra sự tự nhiên và chân thực. Các đóng góp của chúng tôi trong nghiên này bao gồm các nội dung sau:

* Chúng tôi phát triển một phương pháp dựa trên kỹ thuật học sâu kết hợp phương pháp biến đổi Hidden Affine Transformation và kỹ thuật xử lý ảnh để có thể tổng hợp một chuỗi video động từ một bức ảnh tĩnh duy nhất, trong đó tích hợp một cách mượt mà các chuyển động của môi, mắt và các phần có liên quan trên khuôn mặt để tạo ra video từ một hình ảnh tĩnh một cách chân thực và sinh động.
* Chúng tôi sử dụng mô hình học sâu cGAN để học các phép biến đổi Affine, các tham số Affine này được tính toán dựa trên các thông tin từ ảnh đầu vào như: ảnh khuôn mặt và được sử dụng để tạo ra các khung hình sau đó tổng hợp thành video có chuyển động môi, mắt và các phần có liên quan trên khuôn mặt.
* Chúng tôi tiến hành các thực nghiệm và đánh giá đầy đủ trên các tập dữ liệu đa dạng, nhằm chứng minh tính hiệu quả và độ ổn định của mô hình đề xuất của mình.
* Chúng tôi so sánh phương pháp của chúng tôi với các phương pháp hiện có và đánh giá kết quả theo chất lượng hình ảnh, tính thực tế và sự liên kết.

Ngoài phần giới thiệu, phần còn lại của bài báo bao gồm các nội dung sau. Phần 2 trình bày về các nghiên cứu có liên quan. Mô hình đề xuất của chúng tôi được trình bày ở Phần 3. Phần 4 trình bày về thực nghiệm và đánh giá so sánh kết quả của chúng tôi với các phương pháp khác. Phần 5 trình bày về kết luận về nghiên cứu của chúng tôi và hướng phát triển trong tương lai.

**2 CÁC NGHIÊN CỨU LIÊN QUAN**

Tạo video từ hình ảnh là quá trình biến đổi một loạt các ảnh tĩnh thành một chuỗi các khung hình liên tiếp nhằm tạo ra một video chuyển động của các vật thể có trong bức ảnh. Có nhiều phương pháp để giải quyết bài toán này, đa số các phương pháp sử dụng phương pháp học sâu vì tính hiệu quả của phương pháp này.

Sử dụng mô hình cINN và các phương pháp khác như phép biến đổi Affine, học sâu phi tuyến và mô hình autoregressive để tổng hợp video từ các hình ảnh đầu vào [1]. Tuy nhiên phương pháp trên tồn tại một số hạn chế như khả năng giới hạn của mô hình ước lượng , và khó khăn trong việc xử lý các phân bố lớn, phức tạp.

Yang Zhou và cộng sự [2] đã giới thiệu một phương pháp sử dụng học sâu để tạo ra video động của người nói với môi chuyển động mượt mà và diễn đạt hình ảnh chân dung tự nhiên. Mô hình có thể tạo ra hoạt hình biểu diễn sự đồng bộ miệng, biểu cảm khuôn mặt cá nhân và động tác đầu một cách tốt hơn so với công nghệ tiên tiến hiện tại, song mô hình trên vẫn còn tồn tại một số điểm hạn chế không thể sinh ra các hình ảnh người nói chuyện với độ phân giải cao

Guangyao Shen và cộng sự [3] đề xuất một phương pháp để chuyển đổi ảnh khuôn mặt đến video bằng cách sử dụng một phép biến đổi affine ẩn. Phương pháp này được thực hiện bằng cách sử dụng một mô hình học sâu có tên là AffineGAN, với kiến trúc dựa trên mô hình GAN (Generative Adversarial Networks). Phương pháp này có thể tạo ra các video mới với nhiều biểu cảm khác nhau trên khuôn mặt, và tạo ra các video mượt hơn, có độ chân thực cao hơn. Hạn chế của phương pháp này là cần sự chuẩn bị dữ liệu cẩn thận; giới hạn của biến đổi affine; phụ thuộc vào chất lượng dữ liệu đầu vào; khả năng tổng quát hóa và mất mát thông tin.

Haomiao Ni và cộng sự [4] đề xuất mô hình Latent Flow Diffusion Models (LFDMs) là một lớp mô hình sinh dựa trên dòng latent (latent flow), được sử dụng để mô phỏng phân phối dữ liệu phức tạp như ảnh và video. Mô hình đề xuất có thể tạo ra video bằng cách biến đổi hình ảnh cho trước bằng các chuỗi flow được tạo ra trong không gian ẩn dựa trên điều kiện lớp. Hạn chế của mô hình này là LFDM giới hạn trong việc xử lý video chứa một đối tượng di chuyển duy nhất; LFDM hiện tại được điều kiện dựa trên nhãn lớp thay vì mô tả tự nhiên bằng văn bản; so với các mô hình GAN, LFDM chậm hơn rất nhiều khi lấy mẫu với DDPM 1000 bước.

Ming-Yu Liu và cộng sự [5] đề xuất mô hình tạo ra video từ các hình ảnh đơn lẻ. Quá trình này có thể bao gồm việc sinh ra các khung hình trung gian giữa các khung hình có sẵn, để tạo ra dự báo liên tục về chuyển động. Ưu điểm của phương pháp này là cung cấp khả năng tạo ra video động từ các hình ảnh đơn lẻ, mở rộng khả năng tạo ra video từ dữ liệu hình ảnh thưa; tạo ra các khung hình video có tính chân thực và liên tục về chuyển động. Tuy nhiên, phương pháp này tồn tại một số hạn chế như đòi hỏi lượng dữ liệu huấn luyện lớn và tính toán phức tạp; cần tinh chỉnh tham số và quá trình huấn luyện kỹ lưỡng để đạt được chất lượng video tốt; có thể gặp khó khăn trong việc tái tạo các chi tiết và chuyển động tự nhiên trong video.

Long Zhao và cộng sự [6] giới thiệu một phương pháp sinh video từ ảnh đầu vào bằng cách dự đoán và cải tiến chuyển động dư thừa. Mô hình trên có thể dự đoán chuyển động dư thừa, có khả năng tạo ra video chất lượng cao. Điểm hạn chế của phương pháp này là phụ thuộc vào khung hình gốc; độ phức tạp tính toán lớn, giới hạn trong việc xử lý các trường hợp phức tạp như đối tượng di chuyển nhanh, chuyển động phức tạp hoặc sự thay đổi nhanh về ánh sáng trong video.

Dựa trên những nghiên cứu trước đây, trong bài nghiên cứu này chúng tôi đề xuất và sử dụng mô hình hiệu quả để tổng hợp video động từ một bức ảnh tĩnh, tích hợp chuyển động môi, mắt và các phần có liên quan trên khuôn mặt vào video tổng hợp để tạo ra sự tự nhiên và chân thực dựa trên kỹ thuật học sâu kết hợp phương pháp biến đổi Hidden Affine Transformation và kỹ thuật xử lý ảnh. Chúng tôi sử dụng mạng nơ ron tích chập CNN cho việc trích xuất các đặc trưng hình ảnh khuôn mặt và conditional Generative Adversarial Network (cGAN) để tổng hợp video từ các đặc trưng hình ảnh. Mô hình tổng quát sẽ được trình bày trong phần tiếp theo.

**3 MÔ HÌNH ĐỀ XUẤT**

**3.1 Mô hình tổng quát**

Từ tập dữ liệu thô ban đầu chúng tôi sử dụng các phương pháp tiền xử lý dữ liệu sau đó đưa qua mô hình học sâu để tiến hành tổng hợp video từ hình ảnh khuôn mặt. Tiếp đến sẽ lưu lại các tham số của mô hình và tối ưu các tham số đó. Để mô hình cho ra video chuyển động chân thực và sinh động hơn. Mô hình tổng quát của chúng tôi được trình bày trong Hình 1.

Dữ liệu thô

Tiền xử lý   
dữ liệu

Huấn luyện mô hình

AffineGAN

Lưu lại các tham số của mô hình

Hình 1: Sơ đồ mô hình tổng quát

**3.2 Các thành phần của mô hình**

**3.2.1 Conditional Generative Adversarial Network (cGAN)**

Generative Adversarial Networks (cGAN) được đề xuất bởi Mirza và Osindero vào năm 2014[7]. cGAN là một dạng mạng Generative Adversarial Network (GAN) trong lĩnh vực học sâu có khả năng tạo ra dữ liệu mới dựa trên một điều kiện đầu vào cụ thể.

A picture containing screenshot, text, diagram, circle

Description automatically generated

Hình 2: Cấu trúc cGAN [7]

Cấu trúc cGAN gồm 2 phần chính: mạng sinh (generator) và mạng phân biệt (discriminator).

**Generator:**

Mạng sinh nhận đầu vào là một nhãn và một hình ảnh khuôn mặt. Nhãn chứa các thông tin về biểu cảm hoặc các thuộc tính khác của khuôn mặt. Mạng sinh sử dụng kiến trúc dựa trên các lớp convolutional và deconvolutional để ánh xạ từ hình ảnh khuôn mặt đến các khung hình video. Lớp convolutional giúp trích xuất đặc trưng từ ảnh khuôn mặt và phép biến đổi affine ẩn, trong khi các lớp deconvolutional giúp tạo ra các khung video chi tiết.

Chúng tôi sử dụng CNN trong mạng sinh (generator) để phân tích và rút trích các đặc trưng quan trọng từ hình ảnh khuôn mặt, bao gồm các đặc trưng hình dạng và cấu trúc khuôn mặt, các chi tiết như mắt, mũi, miệng và các đặc trưng diễn cảm. Các đặc trưng này sau đó được sử dụng để thực hiện phép biến đổi Hidden Affine và tạo ra video từ hình ảnh khuôn mặt đầu vào.

**Discriminator:**

Mạng phân biệt có nhiệm vụ phân loại các khung hình video là thật hay giả. Khác với mạng phân biệt truyền thống, mạng phân biệt trong cGAN nhận thêm một nhãn làm đầu vào, giúp phân biệt dựa trên cả nội dung hình ảnh và nhãn. Mạng phân biệt nhận đầu vào là khung hình video và nhãn, sau đó đi qua các lớp convolutional để trích xuất đặc trưng.

Chúng tôi sử dụng cGAN để chuyển đổi hình ảnh khuôn mặt thành video thông qua quá trình tạo ra các khung hình tiếp theo trong chuỗi video.

Mô hình cGAN được sử dụng như sau:

**Generative Network:**

Mạng sinh nhận vào một hình ảnh khuôn mặt đầu vào và cố gắng tạo ra video tương ứng với hình ảnh đó. Mô hình cGAN sử dụng mạng sinh có cấu trúc Encoder-Decoder, trong đó:

Encoder làm nhiệm vụ trích xuất các đặc trưng từ hình ảnh khuôn mặt đầu vào.

Sau đó, mạng sinh sử dụng một Hidden Affine Transformation để ánh xạ các đặc trưng từ không gian hình ảnh sang không gian video. Hidden Affine Transformation là một biến đổi tuyến tính bí mật ánh xạ các đặc trưng từ không gian khuôn mặt sang không gian video. Điều này giúp tạo ra sự chuyển động tự nhiên trong video kết quả.

Decoder: Cuối cùng, các đặc trưng được ánh xạ qua mạng decoder để tạo ra video đầu ra.

**Discriminative Network:**

Mạng phân biệt được đào tạo để phân biệt giữa các video được tạo ra bởi mạng sinh và các video thực tế. Mạng phân biệt sử dụng một kiến trúc mạng phân loại thông thường, được huấn luyện để phân biệt đúng giữa video khuôn mặt tạo ra và video thực tế. Nhiệm vụ của mạng sinh là tạo ra các video mà mạng phân biệt không thể phân biệt được với video thực tế.

**3.2.2 Phép biến đổi Hidden Affine**

Phép biến đổi affine ẩn (hidden affine transformation) là một phương pháp để áp dụng các phép biến đổi affine trên dữ liệu học sâu mà không cần xác định rõ ràng các tham số affine. Thay vào đó, mô hình học sâu sẽ học tự động các phép biến đổi affine dựa trên dữ liệu huấn luyện.

Phép biến đổi affine ẩn cho phép mô hình học cách tạo ra các khung hình video từ ảnh khuôn mặt ban đầu bằng cách thay đổi tỷ lệ, xoay và dịch chuyển các đặc trưng trên khuôn mặt.

Cách thực hiện phép biến đổi affine ẩn trong mô hình học sâu thường bao gồm sử dụng mạng nơ-ron để học và dự đoán các tham số affine tương ứng với từng khung hình video. Các tham số affine này được tính toán dựa trên thông tin đầu vào, chẳng hạn như ảnh khuôn mặt, và được sử dụng để tạo ra các khung hình video mới.

Trong bài nghiên cứu này chúng tôi sử dụng phép biến đổi Hidden Affine để cải thiện quá trình biến đổi hình ảnh khuôn mặt thành video bằng cách áp dụng các biến đổi tuyến tính ẩn để tái tạo các biểu cảm và chuyển động tự nhiên trong khuôn mặt.

* 1. **Phương pháp đánh giá**

**3.3.1 Các độ đo và các phương pháp so sánh**

Để đánh giá mô hình một cách chi tiết và định lượng, chúng tôi sử dụng các độ đo sau:

PSNR (Peak Signal-to-Noise Ratio) [8]: Đánh giá chất lượng hình ảnh hoặc video. Độ đo này được sử dụng để đo độ tương đồng giữa hình ảnh (hoặc video) gốc và hình ảnh (hoặc video) tái tạo.

SSIM (Structural Similarity Index Measure) [8]: Đánh giá sự tương đồng cấu trúc giữa hai hình ảnh hoặc video. Đây là độ đo được sử dụng để đo mức độ tương tự về cấu trúc, thông tin và hình dạng giữa hình ảnh (hoặc video) gốc và hình ảnh (hoặc video) tái tạo.

ACD (Average Content Distance) [9]: Đánh giá tính nhất quán về nội dung trong các video được tạo ra. Độ đo này được tính bằng cách lấy trung bình của tất cả các khoảng cách L2 giữa các đặc trưng khuôn mặt của hai khung hình liên tiếp trong một video được tạo ra.

ACD-I (Average Content Distance - Identity) [10] là một phiên bản mở rộng của tiêu chí đánh giá ACD (Average Content Distance) để đo tính bảo tồn định danh của khuôn mặt đầu vào trong video được tạo ra. ACD-I tính trung bình khoảng cách L2 giữa các đặc trưng khuôn mặt trong các khung hình của video và hình ảnh đầu vào. Nó đo lường mức độ bảo tồn định danh của khuôn mặt gốc trong quá trình chuyển đổi thành video. Trong nghiên cứu này, chúng tôi sử dụng OpenFace[11], một mô hình sâu được huấn luyện để nhận dạng khuôn mặt và có khả năng vượt qua hiệu suất của con người, để trích xuất các đặc trưng khuôn mặt.

Các điểm số SSIM và PSNR càng cao thì chất lượng video được tạo ra càng tốt, các điểm số ACD thấp hơn cho thấy sự tương tự giữa các khuôn mặt trong các khung video liên tiếp được tạo ra, và các giá trị ACD-I thấp hơn cho thấy sự tương đồng cao hơn giữa các khuôn mặt trong ảnh đầu vào và các video được tạo ra.

Để đánh giá mô hình của chúng tôi với các mô hình khác, chúng tôi lựa chọn 5 mô hình tiến tiến, đạt kết quả tốt cho bài toán này được xây dựng dựa trên mô hình GAN để so sánh đó là: VGAN[12], MoCoGAN[13], GANimation[14], ImaGINator[15], và CwGAN[16].

**3.3.2 Hàm mất mát**

**MSE (Mean Square Error)** [17]:

Được sử dụng trong LSGAN (Least Squares GAN).

Hàm mất mát này tính toán sai số trung bình giữa đầu ra của Discriminator và nhãn mục tiêu.

Thích hợp cho bài toán GAN khi muốn tối thiểu hóa sai số bình phương giữa đầu ra và nhãn.

**BCE (Binary Cross-Entropy)** [18]:

Một hàm mất mát (loss function) phổ biến trong các bài toán phân loại nhị phân trong Machine Learning. Hàm BCE được sử dụng để đo lường sai khác giữa các dự đoán nhị phân và nhãn thực tế trong bài toán phân loại nhị phân. Thông thường, đầu ra của một mô hình phân loại nhị phân được ánh xạ vào khoảng giá trị [0, 1], ví dụ như sử dụng hàm kích hoạt Sigmoid. BCE tính toán sai số chéo (cross-entropy) giữa dự đoán nhị phân và nhãn thực tế dưới dạng một giá trị số.

Hàm mất mát này tính toán cross entropy (mất mát entropy chéo) giữa đầu ra của Discriminator và nhãn mục tiêu. Thích hợp cho bài toán GAN khi muốn tối thiểu hóa sự khác biệt giữa phân bố xác suất của đầu ra và nhãn.

1. **THỰC NGHIỆM**

**4.1 Dữ liệu**

Trong bài báo này chúng tôi sử dụng tập dữ liệu CK-Mixed. CK-Mixed là bộ dữ liệu Cohn-Kanade (CK+)[19] phổ biến ở phân tích biểu cảm khuôn mặt, trong đó mỗi biểu cảm được mở ra từ trạng thái trung lập ban đầu một cách đơn điệu. Bộ dữ liệu này gồm 593 video của 6 loại cảm xúc, trong đó hầu hết các video đều ở thang độ xám. Sau khi tiến hành xử lý dữ liệu video chúng tôi trích xuất các khung hình từ video thu được tập dữ liệu ảnh gồm 8000 ảnh. Chúng chúng tôi tiến hành chia tập dữ liệu theo tỉ lệ 8:2. Bảng 1 trình bày khái quát về Bộ dữ liệu CK-Mixed, một số hình ảnh về bộ dữ liệu được trình bày trong Hình 3.

Bảng 1: Bộ dữ liệu

|  |  |
| --- | --- |
| Datasets | Số lượng |
| Train | 5400 ảnh |
| Test | 1600 ảnh |

A group of people posing for a photo

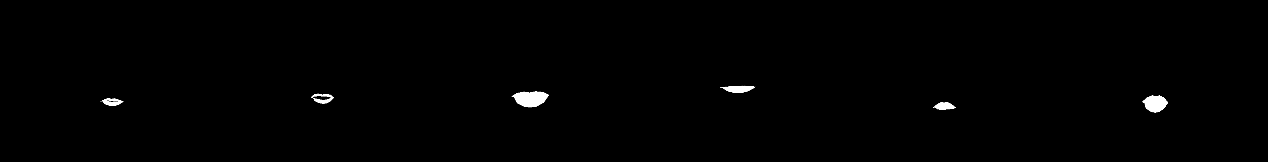
Description automatically generated with low confidence

Hình 3: Một số hình ảnh trong tập dữ liệu CK-Mixed

**4.1.1 Dlib**

Để xác định vị trí phần môi trên khuôn mặt trong ảnh chúng tôi sử dụng Dlib [20] là một thư viện mã nguồn mở được phát triển bởi Davis E. King.

Sau khi xác định được vị trí môi, chúng tôi áp dụng thư viện OpenCV [21] để cắt phần môi đó ra từ ảnh gốc, sau đó tiến hành chuyển đổi phần môi này sang ảnh nhị phân để tạo patch dùng để vá lỗi cho video khi chuyển động. Hình 4 mô tả chi tiết patch phần môi sau khi được xử lý của chúng tôi.



Hình 4: Patch phần môi

**4.2 Kết quả**

Dựa trên phương pháp đề xuất trên, chúng tôi sẽ tạo video từ một hình ảnh đầu vào duy nhất. Phương pháp này có khả năng tạo ra video chuyển động mượt mà và chân thực từ hình ảnh khuôn mặt đầu vào. Các kết quả đã được đánh giá dựa trên các tiêu chí như sự tương đồng với hình ảnh gốc, sự chân thực và tính đa dạng của video kết quả. Hình 5 trình bày một ví dụ về kết quả dự đoán sau khi thực hiện mô hình.

A collage of a person

Description automatically generated with medium confidence

Hình 5: Kết quả dự đoán sau khi thực hiện mô hình

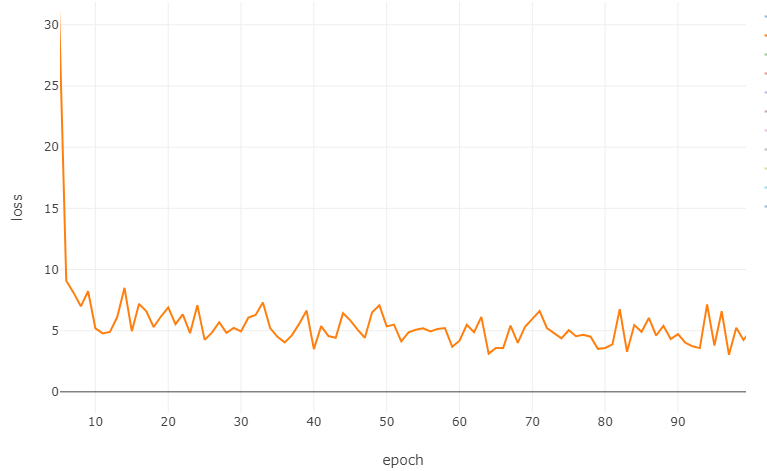
Như phần trước chúng tôi đã trình bày, chúng tôi lựa chọn 5 mô hình so sánh đó là: VGAN[12], MoCoGAN[13], GANimation[14], ImaGINator[15], và CwGAN[16].

Về các mô hình dùng để so sánh chúng tôi đã thực nghiệm lại trên cơ sở mã nguồn công khai của VGAN và ImaGINAtor do các tác giả cung cấp. Để áp dụng GANimation, chúng tôi căn chỉnh khuôn mặt và sau đó trích xuất các đặc điểm biểu hiện (AUs) thông qua OpenFace [10]. Trong giai đoạn dự đoán, chúng tôi chọn AU biểu hiện phong phú nhất từ tập dữ liệu huấn luyện làm nhãn cho mỗi biểu cảm mục tiêu. Đối với MoCoGAN chúng tôi sử dụng kết quả được đưa ra bởi Wang et al. (2020b)[15]. Đối với CwGAN chúng tôi sử dụng kết quả trong bài báo gốc.

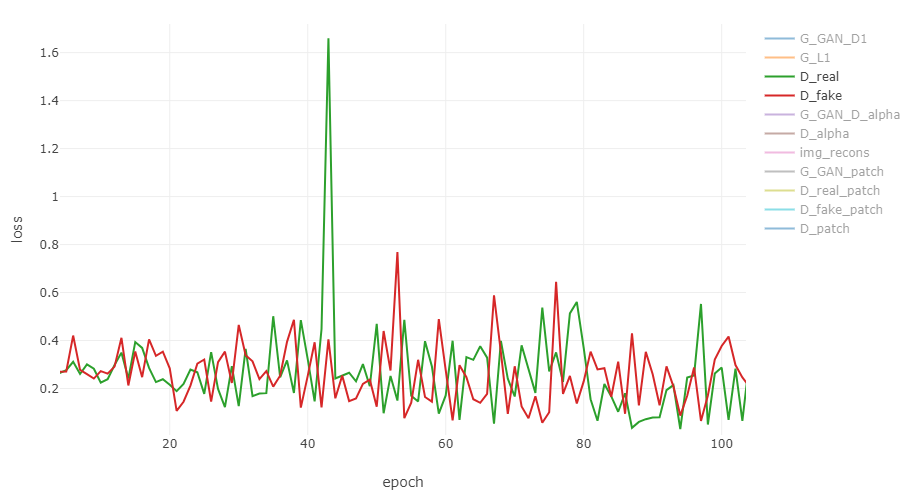
Kết quả so sánh giữa 5 phương pháp với mô hình đề xuất của chúng tôi được trình bày trong Bảng 2. Kết quả về hàm mất mát đánh giá sự khác biệt giữa ảnh được tạo (fake\_image) và ảnh thuật (real\_image) được trình bày trong Hình 6 và Hình 7 trình bày về kết quả của Discriminator Loss.

Bảng 2: Bảng so sánh giữa các phương pháp

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phương pháp** | **PSNR** | **SSIM** | **ACD** | **ACD-I** |
| VGAN | 16.32 | 0.41 | 0.14 | 1.55 |
| MoCoGAN | 18.16 | 0.58 | 0.15 | 0.90 |
| GANimation | 26.54 | 0.89 | 0.10 | 0.35 |
| ImaGINator | 20.29 | 0.85 | 0.08 | 0.29 |
| CwGAN | 25.90 | 0.90 | 0.11 | 0.120 |
| **AffineGAN** | **35.50** | **0.91** | **0.06** | **0.16** |

****

Hình 6: Hàm mất mát đánh giá sự khác biệt giữa ảnh được tạo (fake image) và ảnh thuật (real image)



Hình 7: Discriminator Loss

Nhìn vào kết quả trong Bảng 2, chúng ta thấy được rằng phương pháp đề xuất của chúng tôi đạt kết quả tốt nhất so với các phương pháp so sánh khác ở cả 4 độ đo PSNR, SSIM, ACD và ACD-I. Kết quả này cho thấy phương pháp học sâu cGAN kết hợp với Hidden Affine Transformation cho kết quả tốt hơn so với các phương pháp GAN khác.

Chúng tôi đánh giá sự khác biệt giữa ảnh được tạo (`fake\_image`) và ảnh thật (`real\_image`) dựa trên L1Loss [22]. Mục tiêu là giảm thiểu độ sai lệch giữa ảnh được tạo và ảnh thật. Ở Hình 6, ta thấy loss giảm dần theo thời gian khi Generator cải thiện khả năng tạo ảnh. Mục tiêu cuối cùng là đạt được giá trị G\_L1 thấp nhất, khi Generator tạo ra các ảnh có sự tương đồng cao so với ảnh thật, đồng nghĩa với việc tạo ra các ảnh chất lượng cao và khó phân biệt với ảnh thật.

Đối với Hình 7, D\_real và D\_fake là kết quả đầu ra của bộ phân biệt (Discriminator) khi đưa vào dữ liệu thật (real) và dữ liệu giả tạo (fake) tương ứng. D\_real là kết quả đầu ra của bộ phân biệt khi đưa vào dữ liệu thật. D\_fake là kết quả đầu ra của bộ phân biệt khi đưa vào dữ liệu giả tạo. Mục tiêu của mô hình là tạo ra dữ liệu giả tạo (fake) mà bộ phân biệt khó phân biệt được với dữ liệu thật (real). Bằng cách tối ưu hóa hàm mất mát, mô hình cố gắng làm cho D\_real gần với 1 và D\_fake gần với 0, từ đó tạo ra dữ liệu giả tạo chất lượng cao.

Tuy nhiên, khi phân tích chi tiết chúng tôi thấy rằng mô hình đề xuất của chúng tôi cũng có những hạn chế như: vẫn còn một số trường hợp đặc biệt khi tổng hợp video gặp khó khăn, ví dụ như khi bức ảnh ban đầu có chất lượng thấp hoặc không rõ ràng, hoặc khi các chuyển động trên khuôn mặt rất phức tạp và tinh tế. Các hạn chế này có thể gây ra các vấn đề như nhiễu hình ảnh, video không tự nhiên hoặc video không chính xác.

1. **KẾT LUẬN**

Kết quả thực nghiệm của nghiên cứu đã chứng minh tính hiệu quả và tiềm năng của phương pháp đề xuất trong tạo ra video động từ một bức ảnh duy nhất. Bằng cách sử dụng mô hình CNN để trích xuất đặc trưng và mô hình cGAN để tổng hợp video, phương pháp đã đạt được kết quả ấn tượng. Các video tổng hợp có chứa chuyển động tự nhiên và chân thực của môi, mắt và các phần khác trên khuôn mặt. Thực nghiệm trên các tập dữ liệu đa dạng đã chứng minh tính ổn định và khả năng tái tạo tốt của phương pháp trong nhiều trường hợp khác nhau.

Để cải thiện và phát triển phương pháp trong tương lai, chúng tôi có đề xuất một số hướng tiếp cận để có thể tiếp tục nghiên cứu chuyên sâu hơn. Đầu tiên, mở rộng kiến trúc mô hình để xử lý các trường hợp đặc biệt và khó khăn hơn, bao gồm chất lượng hình ảnh thấp, đa dạng chuyển động, và sự biến đổi khuôn mặt lớn. Thứ hai, nghiên cứu có thể tập trung vào cải thiện tính thực tế của video tổng hợp, đảm bảo rằng các chuyển động và chi tiết trên khuôn mặt được phản ánh chính xác. Thứ ba, khám phá khả năng ứng dụng của phương pháp trong các lĩnh vực khác nhau, chẳng hạn như điện ảnh, truyền thông và giải trí.

Với sự phát triển tiếp tục và nghiên cứu sâu hơn, phương pháp đề xuất có thể trở thành một công cụ quan trọng và mạnh mẽ trong việc tạo ra video động từ một bức ảnh duy nhất, với nhiều ứng dụng tiềm năng trong nhiều lĩnh vực khác nhau.

**TÀI LIỆU HAM KHẢO**

1. M. Dorkenwald, T. Milbich, A. Blattmann, R. Rombach, K. G. Derpanis and B. Ommer. (2021) Stochastic Image-to-Video Synthesis using cINNs.[Online]. Available:<https://arxiv.org/pdf/2105.04551.pdf>
2. Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis And D. Li. (2021) MakeItTalk: Speaker-Aware Talking-Head Animation. [Online]. Available:<https://arxiv.org/pdf/2004.12992.pdf>
3. G. Shen, W. Huang, C. Gan, M. Tan, J. Huang, W. Zhu, and B. Gong. (2019) Facial Image-to-Video Translation by a Hidden Aine Transformation. [Online]. Available:<https://doi.org/10.1145/3343031>
4. H. Ni, C. Shi, K. Li, S. X. Huang and M. R. Min. (2023) Conditional Image-to-Video Generation with Latent Flow Diffusion Models. [Online]. Available:<https://arxiv.org/pdf/2303.13744v1.pdf>
5. M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang and A. Mallya. (2020) Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications. [Online]. Available:<https://arxiv.org/pdf/2008.02793.pdf>
6. L. Zhao, X. Peng, Y. Tian, M. Kapadia and D. Metaxas. (2018) Learning to Forecast and Refine Residual Motion for Image-to-Video Generation. [Online]. Available:<https://arxiv.org/pdf/1807.09951v1.pdf>
7. M. Mirza and S. Osindero. (2014) Conditional Generative Adversarial Nets. [Online]. Available: <https://arxiv.org/pdf/1411.1784.pdf>
8. A. Horé and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 2010, pp. 2366-2369, doi: 10.1109/ICPR.2010.579.
9. S. Tulyakov, M.-Y. Liu, X. Yang, J. Kautz. “Mocogan: Decomposing motion ancontent for video generation”, 2018. Available:<https://ieeexplore.ieee.org/document/8578263>
10. L. Zhao, X. Peng, Y. Tian, M. Kapadia, D. Metaxas. “Learning to forecast and refine residual motion for image-to-video generation”, 2018. Available: <https://arxiv.org/abs/1807.09951v1>
11. B. Amos, B. Ludwiczuk, M. Satyanarayanan. “OpenFace: A general-purpose face recognition library with mobile applications”, 2016. Available:<http://cmusatyalab.github.io/openface/>
12. Vondrick, Carl, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics.Advances in neural information processing systems 29. 2016
13. Tulyakov, Sergey, et al. "Mocogan: Decomposing motion and content for video generation.Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
14. Pumarola, Albert, et al. "Ganimation: Anatomically-aware facial animation from a single image.Proceedings of the European conference on computer vision (ECCV). 2018.
15. Wang, Yaohui, et al. "Imaginator: Conditional spatio-temporal gan for video generation.Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.
16. Otberdout, Naima, et al. Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.2 (2020): 848-863.
17. Schluchter, Mark D. "Mean square error." Encyclopedia of Biostatistics 5 (2005).
18. Li, Li, Miloš Doroslovački, and Murray H. Loew. "Approximating the gradient of cross-entropy loss function." IEEE Access 8 (2020): 111626-111635.
19. Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." 2010 ieee computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010.
20. DLIB: <http://dlib.net/>
21. OpenCV: <https://opencv.org/>
22. Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)." *Geoscientific model development discussions* 7.1 (2014): 1525-1534.