

SINH BIỂU CẢM KHUÔN MẶT DỰA TRÊN GIỌNG NÓI

Trần Hoàng Tuấn¹, Nguyễn Quang Hùng²

¹Học viên thực hiện
Khoa học máy tính, đại học Bách Khoa thành phố Hồ Chí Minh

²Tiến sĩ hướng dẫn khoa học
Khoa học máy tính, đại học Bách Khoa thành phố Hồ Chí Minh

Bảo vệ đề cương luận văn, 01/2021

1 Giới thiệu

2 Các công trình nghiên cứu có liên quan

- Mạng GANs
- Tổng quan tình hình nghiên cứu
- Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]

3 Phương pháp đề xuất

- Ý tưởng thực hiện luận văn
- Tiền xử lý dữ liệu
- Các tập dữ liệu được sử dụng
- Các độ đo được sử dụng
- Môi trường thí nghiệm
- Thực hiện thí nghiệm

4 Kết quả thí nghiệm

5 Kết luận

1 Giới thiệu

2 Các công trình nghiên cứu có liên quan

- Mạng GANs
- Tổng quan tình hình nghiên cứu
- Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]

3 Phương pháp đề xuất

- Ý tưởng thực hiện luận văn
- Tiền xử lý dữ liệu
- Các tập dữ liệu được sử dụng
- Các độ đo được sử dụng
- Môi trường thí nghiệm
- Thực hiện thí nghiệm

4 Kết quả thí nghiệm

5 Kết luận

Giới thiệu

Định nghĩa bài toán

- Là một bài toán tạo sinh dữ liệu dạng hình ảnh dựa trên các dạng dữ liệu khác.
- Cho trước 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 người đó đang nói đoạn tiếng nói đã cho một cách chân thực.



Hình 1: Ví dụ về mô hình tạo sinh khuôn mặt

Giới thiệu

Lý do chọn đề tài

- Là nhu cầu cần thiết trong ngành giải trí, phim ảnh, hoạt hình, giúp giảm chi phí sản xuất phim
 - ▶ Phần hóa trang có thể được cắt bớt
 - ▶ Phần kĩ xảo có thể được đơn giản hóa
- Tạo sinh gương mặt đại diện trong trường hợp người nói không muốn lộ diện
- Tạo sinh biên tập viên ảo trong chương trình thời sự, dự báo thời tiết
- Giả lập trợ lý ảo có hình dáng con người

Giới thiệu

Thách thức

- Đây là một đề tài mới lạ, vấn đề tạo sinh dữ liệu chỉ vừa được bùng nổ từ năm 2014 khi mạng GANs xuất hiện
- Tạo sinh dữ liệu cũng là một đề tài khó và phức tạp
- Tạo sinh video từ những dạng dữ liệu khác (hình ảnh, âm thanh) càng làm cho bài toán trở nên thách thức hơn
- Bài toán cũng yêu cầu sức mạnh tính toán lớn và khôi lượng dữ liệu lớn

Giới thiệu

- ① *Mục tiêu:* Xây dựng mô hình có khả năng tạo sinh hình ảnh khuôn mặt người một cách tự nhiên, chính xác.
- ② *Giới hạn:* Tạo sinh hình ảnh trong vùng mặt người. Dữ liệu mẫu được cung cấp ban đầu phải là hình ảnh rõ ràng của khuôn mặt người và một đoạn âm thanh bất kỳ thu âm tiếng nói.
- ③ *Đối tượng:* Các phương pháp mô hình hóa bài toán, học máy, học sâu, mạng GANs và các phương pháp tạo sinh dữ liệu từ mạng GANs, các phương pháp kết hợp đặc trưng hình ảnh, âm thanh để tạo sinh dữ liệu mới.

1 Giới thiệu

2 Các công trình nghiên cứu có liên quan

- Mạng GANs
- Tổng quan tình hình nghiên cứu
- Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]

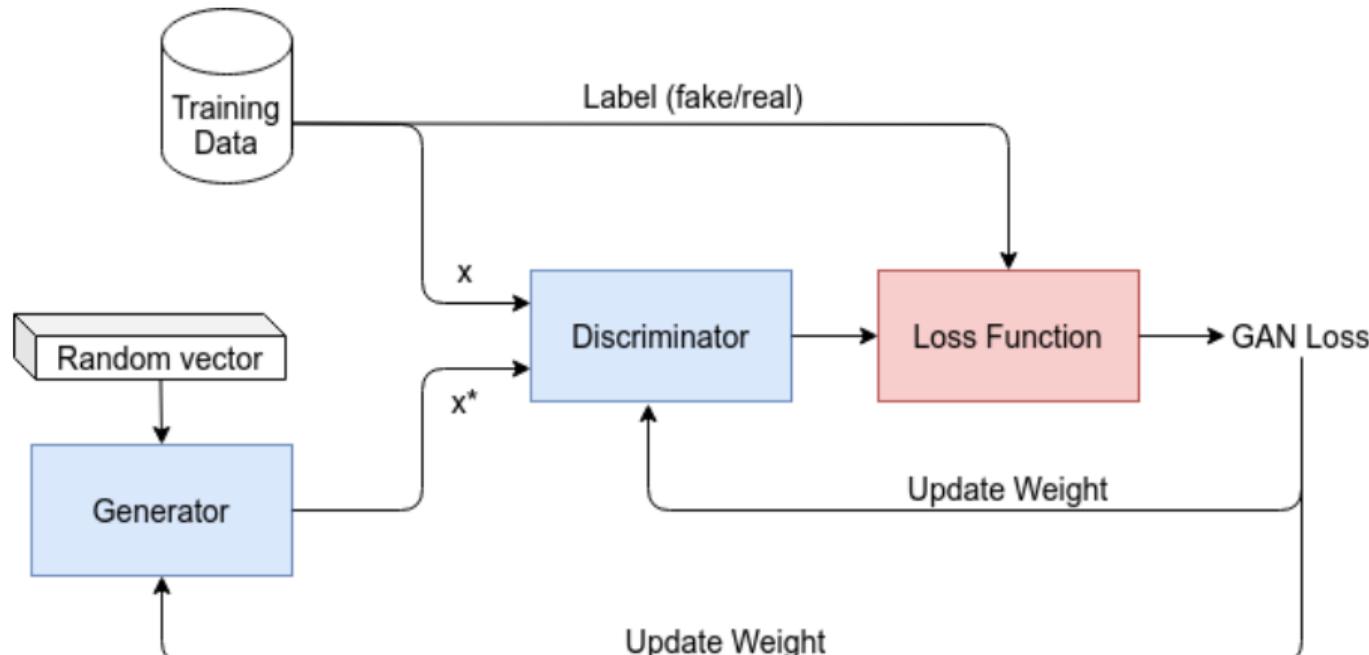
3 Phương pháp đề xuất

- Ý tưởng thực hiện luận văn
- Tiền xử lý dữ liệu
- Các tập dữ liệu được sử dụng
- Các độ đo được sử dụng
- Môi trường thí nghiệm
- Thực hiện thí nghiệm

4 Kết quả thí nghiệm

5 Kết luận

Mạng GANs



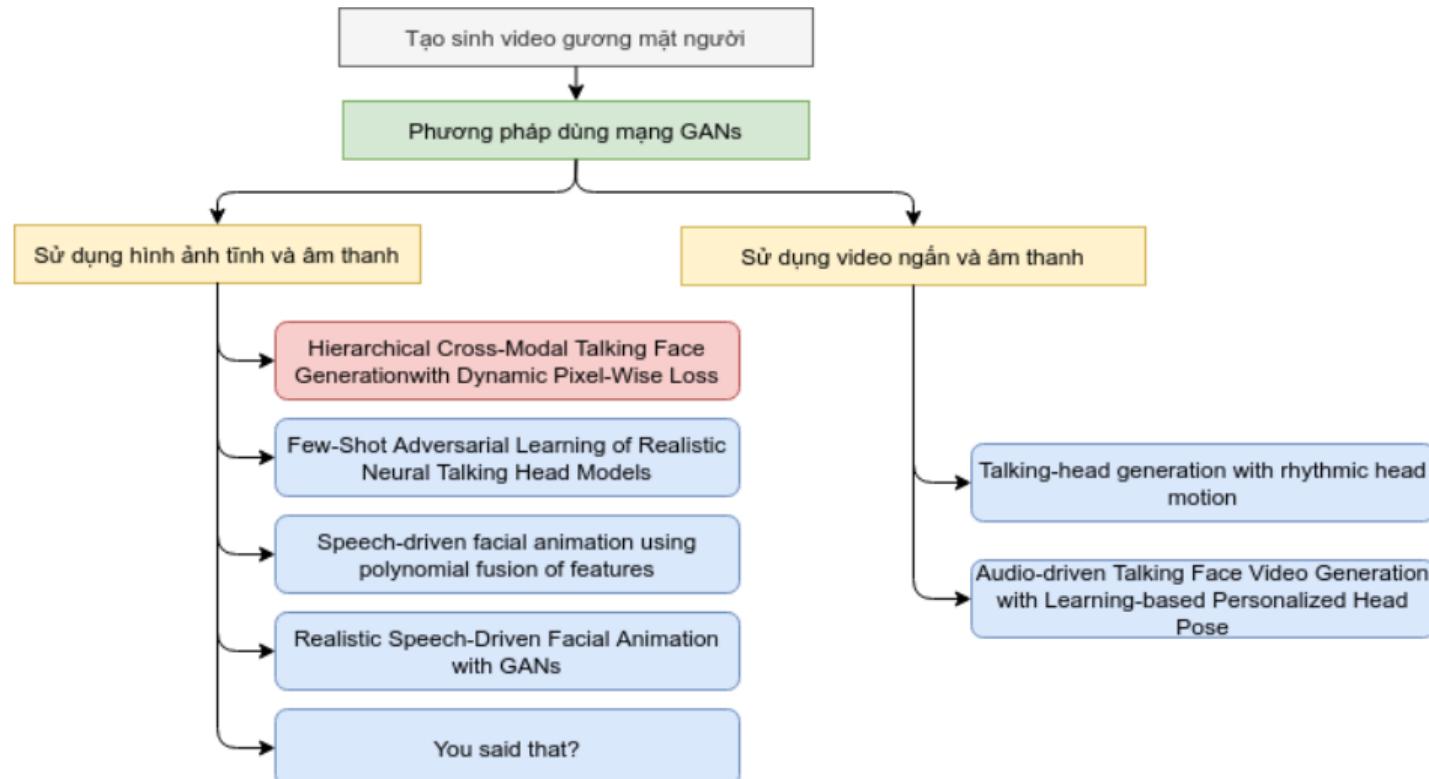
Hình 2: Cấu trúc mạng GANs cơ bản

Mạng GANs



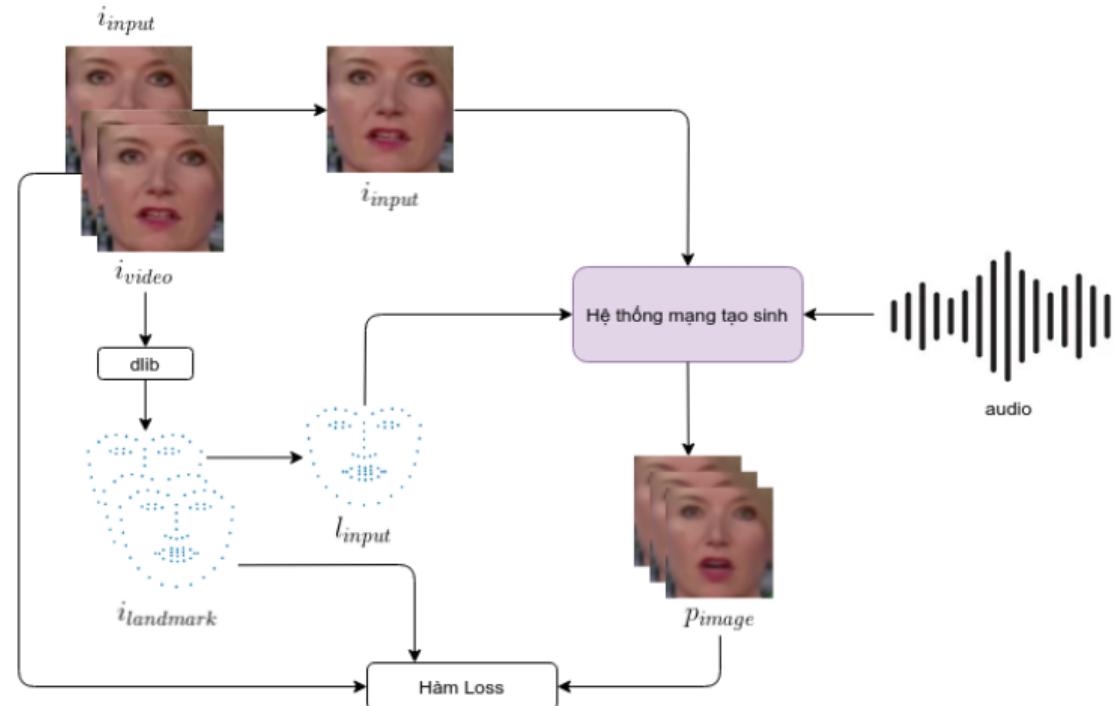
Hình 3: Tạo sinh hình ảnh mặt người bằng mạng GANs theo các năm

Tổng quan tình hình nghiên cứu



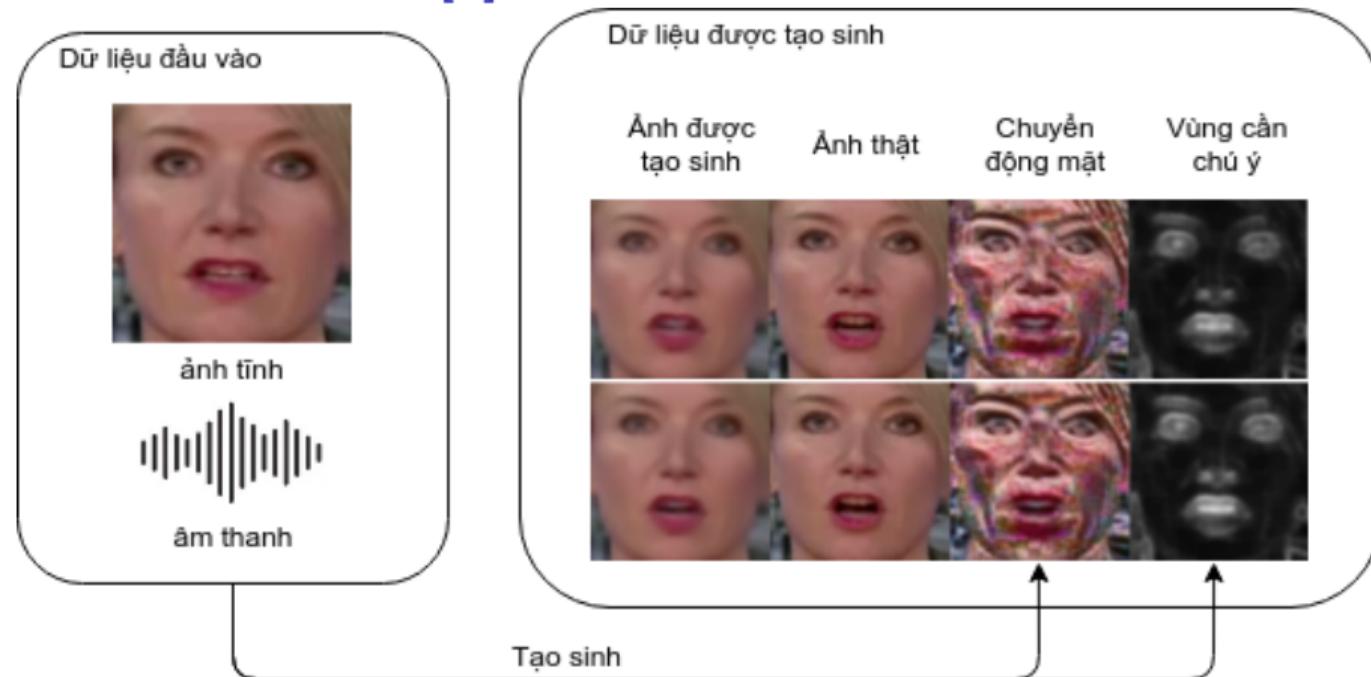
Hình 4: Tổng quan tình hình nghiên cứu

Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]



Hình 5: Phương thức tạo sinh hình ảnh

Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]



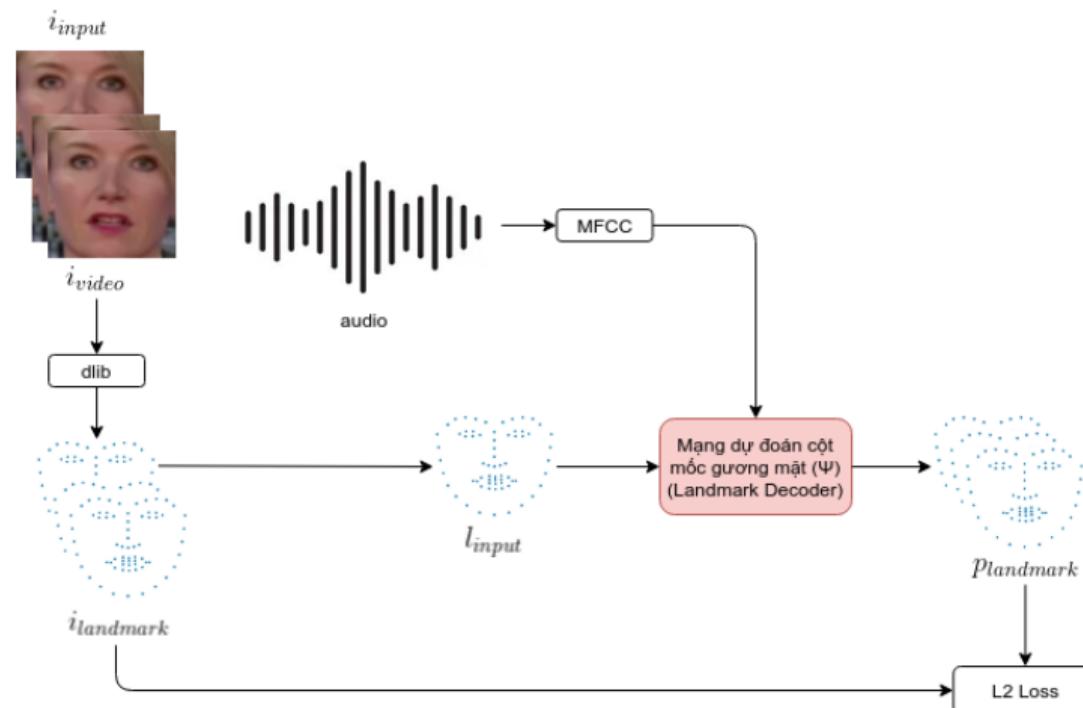
Hình 6: Ý tưởng tạo sinh hình ảnh từ ảnh gốc

Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]

Hệ thống gồm có hai thành phần chính:

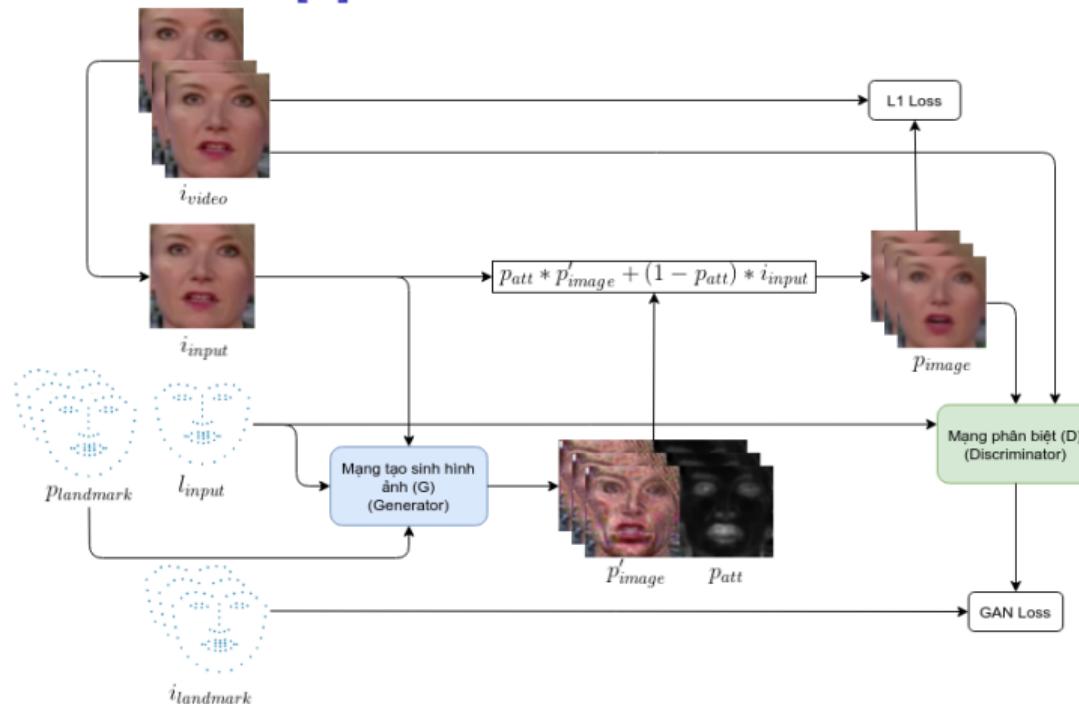
- Mạng dự đoán cột mốc gương mặt (Ψ)
- Hệ thống mạng GANs:
 - ▶ Mạng tạo sinh video (G)
 - ▶ Mạng phân biệt (D)

Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]



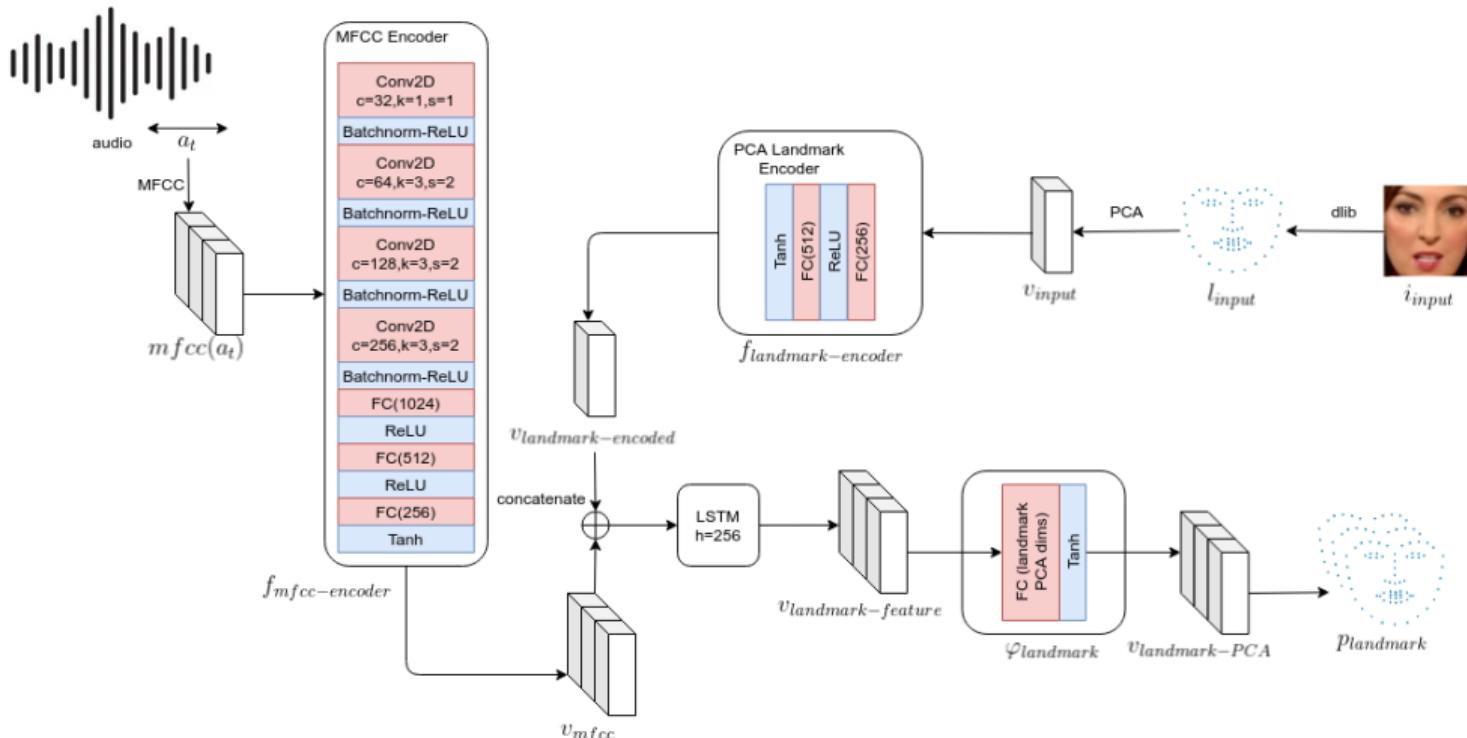
Hình 7: Mạng dự đoán cột mốc gương mặt

Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]



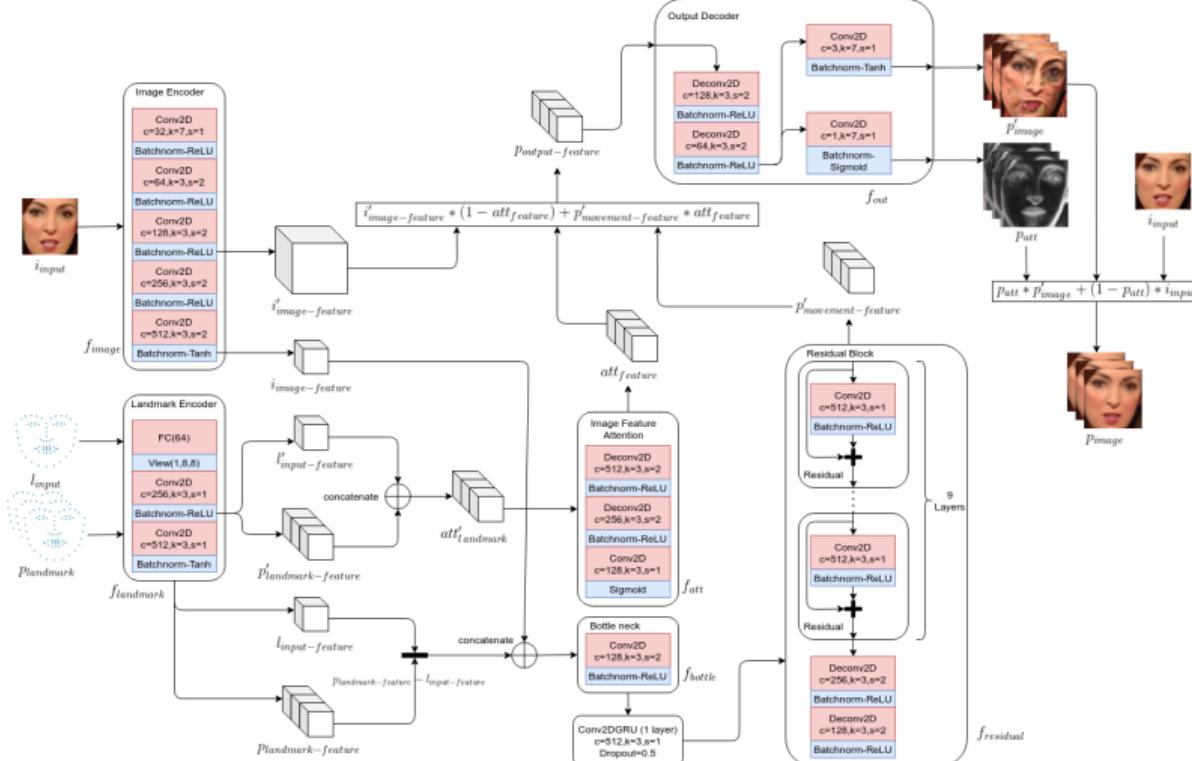
Hình 8: Mạng GANs

Cấu trúc bộ dự đoán cột mốc gương mặt (Landmark Decoder)



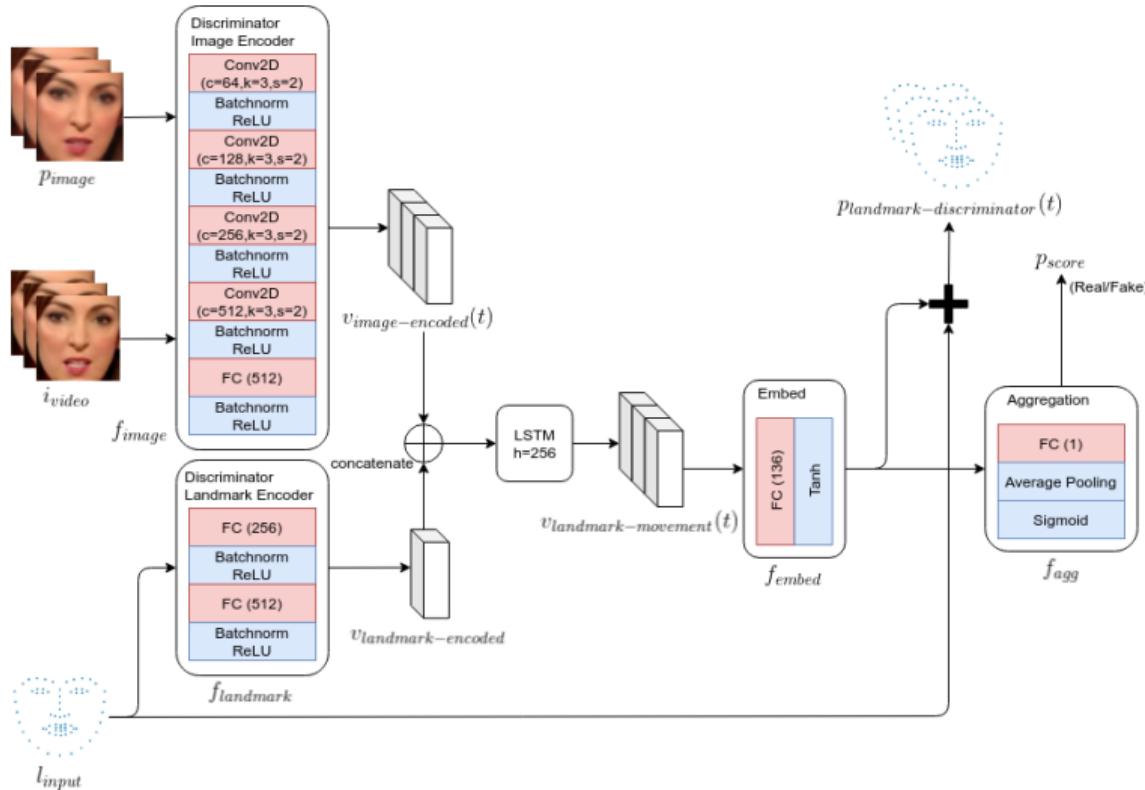
Hình 9: Cấu trúc bộ dự đoán cột mốc gương mặt (Landmark Decoder)

Cấu trúc bộ tạo sinh hình ảnh (Generator)



Hình 10: Cấu trúc bộ tạo sinh hình ảnh (Generator)

Cấu trúc bộ phân biệt (Discriminator)



Hình 11: Cấu trúc bộ phân biệt (Discriminator)

Hàm mất mát tổng

$$\begin{aligned}\mathcal{L}_{gans} &= \mathcal{L}_{gans-dis} + \mathcal{L}_{gans-landmark} \\ \mathcal{L} &= \mathcal{L}_{gans} + \lambda * \mathcal{L}_{pix}\end{aligned}\tag{1}$$

Bằng thực nghiệm ta chọn $\lambda = 10$ để cân bằng giữa \mathcal{L}_{gans} và \mathcal{L}_{pix}

Hàm mất mát cho từng pixel

$$\mathcal{L}_{pix} = \frac{1}{T} \sum_{t=1}^T \|(i_{video}(t) - p_{image}(t)) * (p_{att}(t) + \beta)\|_1 \quad (2)$$

Trong đó:

- $i_{video}(t)$: khung hình tại thời điểm t của video gốc
- $p_{image}(t)$: hình ảnh được tạo sinh tại thời điểm t
- $p_{att}(t)$: mặt nạ chú ý được dự đoán tại thời điểm t
- β : một hằng số để đảm bảo tất cả các điểm trên ảnh đều được học, bằng thực nghiệm ta chọn $\beta = 0.5$

Hàm mất mát GANs cho bộ phân biệt

$$\begin{aligned}\mathcal{L}_{gans-dis} = & \mathbb{E}_{I_{input}, i_{video}} [\log D_s(I_{input}, i_{video})] \\ & + \mathbb{E}_{I_{input}, i_{input}, p_{landmark}} [\log(1 - D_s(I_{input}, G(I_{input}, i_{input}, p_{landmark})))]\end{aligned}\tag{3}$$

Trong đó:

- $I_{input}, i_{input}, p_{landmark}, i_{video}$: đã được giải thích ở các phần trên
- D_s : Mạng phân biệt với đầu ra là xác suất ảnh là ảnh thật
- G : Mạng tạo sinh hình ảnh

Hàm mất mát GANs cho bộ phân biệt

$$\begin{aligned}\mathcal{L}_{gans-landmark} = & \|(D_I(I_{input}, G(I_{input}, i_{input}, p_{landmark})) - I_{orig}) * M_I\|_2^2 \\ & + \|(D_I(I_{input}, i_{video}) - I_{orig}) * M_I\|_2^2\end{aligned}\quad (4)$$

Trong đó:

- $I_{input}, i_{input}, p_{landmark}, i_{video}$: đã được giải thích ở các phần trên
- D_I : Mạng phân biệt với đầu ra là cột mốc gương mặt trên ảnh được tạo sinh
- G : Mạng tạo sinh hình ảnh
- $I_{original}$: Cột mốc gương mặt được trích xuất nguyên gốc từ video i_{video}
- M_I : mặt nạ nhầm chú ý nhiều hơn vào cột mốc gương mặt vùng miệng, theo đó, sai số của cột mốc ở vùng miệng có hệ số 100, trong khi các vùng khác là 1.

Tiền xử lý dữ liệu

Dữ liệu huấn luyện là các video với mặt người đang nói.

- Âm thanh: được lấy mẫu ở tần số thấp hơn và trích xuất đặc trưng MFCC trước khi đưa vào mạng tạo sinh cột mốc
- Hình ảnh: trích xuất và chuẩn hóa hình ảnh gương mặt sao cho mắt, mũi và miệng người nói trong các video có vị trí gần như nhau
- Cột mốc gương mặt: trích xuất và chuẩn hóa cột mốc gương mặt từ các video sao cho tất cả các cột mốc gương mặt đều đồng nhất với nhau

1 Giới thiệu

2 Các công trình nghiên cứu có liên quan

- Mạng GANs
- Tổng quan tình hình nghiên cứu
- Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]

3 Phương pháp đề xuất

- Ý tưởng thực hiện luận văn
- Tiền xử lý dữ liệu
- Các tập dữ liệu được sử dụng
- Các độ đo được sử dụng
- Môi trường thí nghiệm
- Thực hiện thí nghiệm

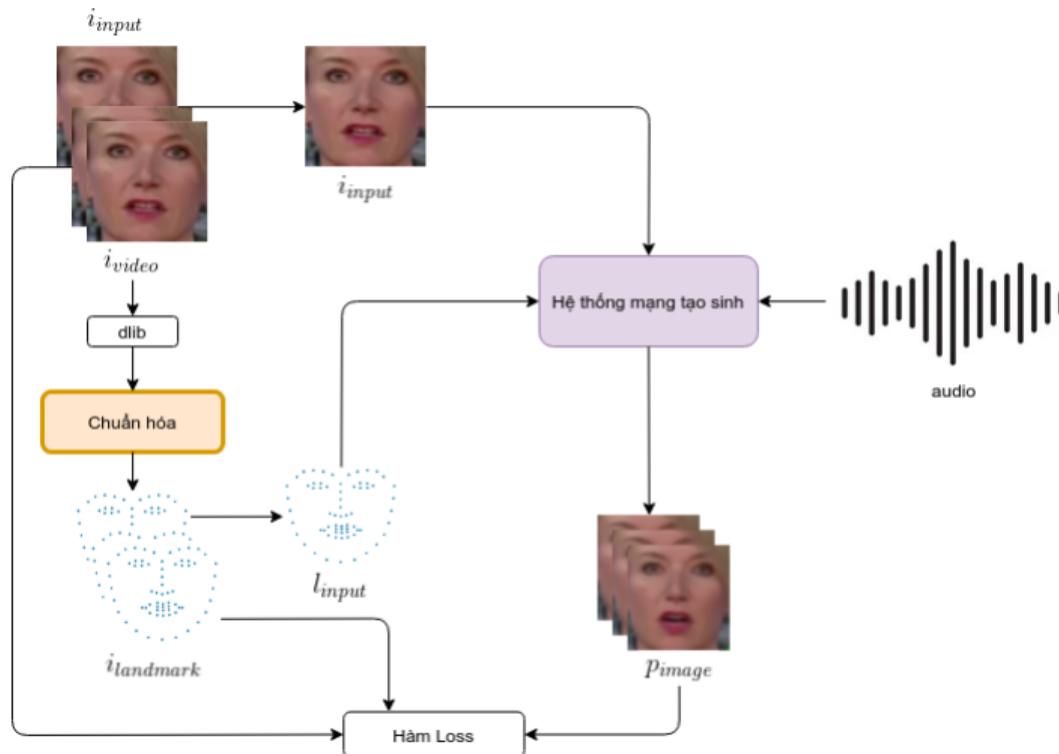
4 Kết quả thí nghiệm

5 Kết luận

Ý tưởng thực hiện luận văn

- Luận văn kế thừa kết quả nghiên cứu của Lele Chen và cộng sự [1]
 - ▶ Nắm rõ ý tưởng của tác giả và quy trình công nghệ
 - ▶ Thực nghiệm lại mô hình tạo sinh hình ảnh của tác giả
- Luận văn cũng kết hợp có chỉnh sửa cách chuẩn hóa cột mốc gương mặt từ nghiên cứu Generating Talking Face Landmarks from Speech[2] để cho kết quả tạo sinh tốt hơn
 - ▶ Cột mốc gương mặt phải có cùng kích thước
 - ▶ Vị trí của mắt, mũi, miệng trên cột mốc gương mặt cần phải được chuyển về một vị trí đồng nhất cho mọi khung hình
 - ▶ Hạn chế việc ảnh hưởng của đặc điểm nhận dạng của khuôn mặt lên cột mốc gương mặt

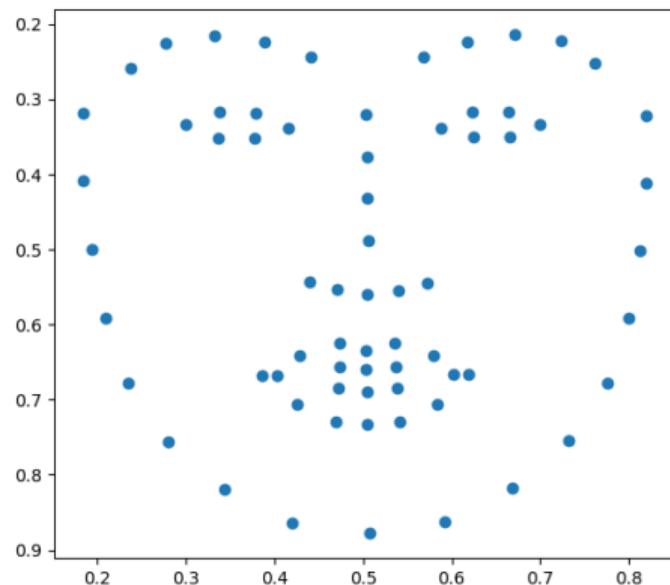
Ý tưởng thực hiện luận văn



Hình 12: Chuẩn hóa cột mốc gương mặt trước khi đưa vào mạng

Tiền xử lý dữ liệu

Ta mong muốn đầu vào của mạng là một cột mốc gương mặt chung nhất và không bị ảnh hưởng bởi đặc điểm mặt người trên video. Vì vậy, ta cần tạo ra cột mốc gương mặt chuẩn bằng cách lấy trung bình cộng của nhiều cột mốc trong nhiều video khác nhau



Hình 13: Cột mốc gương mặt chuẩn

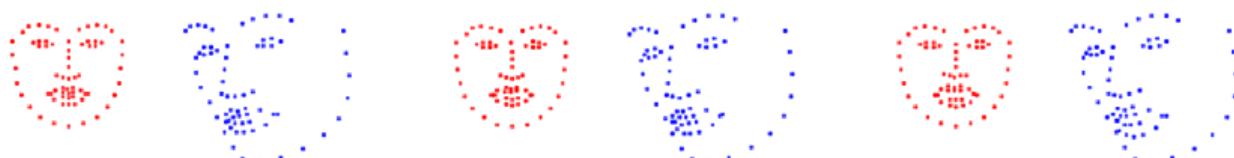
Tiền xử lý dữ liệu

Ý tưởng chuẩn hóa cột mốc gương mặt

- Chuyển vùng mắt của cột mốc gương mặt trong video về vị trí tương ứng trên cột mốc chuẩn
- Tìm đạo hàm của chuỗi cột mốc gương mặt, tính tổng cộng dồn của đạo hàm này để tìm ra sự thay đổi của cột mốc tại thời điểm t so với thời điểm ban đầu
- Chuyển sự thay đổi này lên cột mốc gương mặt chuẩn
- Điều chỉnh lại vùng mũi, miệng của cột mốc gương mặt vừa được tính toán ra bằng các phép biến đổi Affine để khớp hơn với cột mốc chuẩn

Tiền xử lý dữ liệu

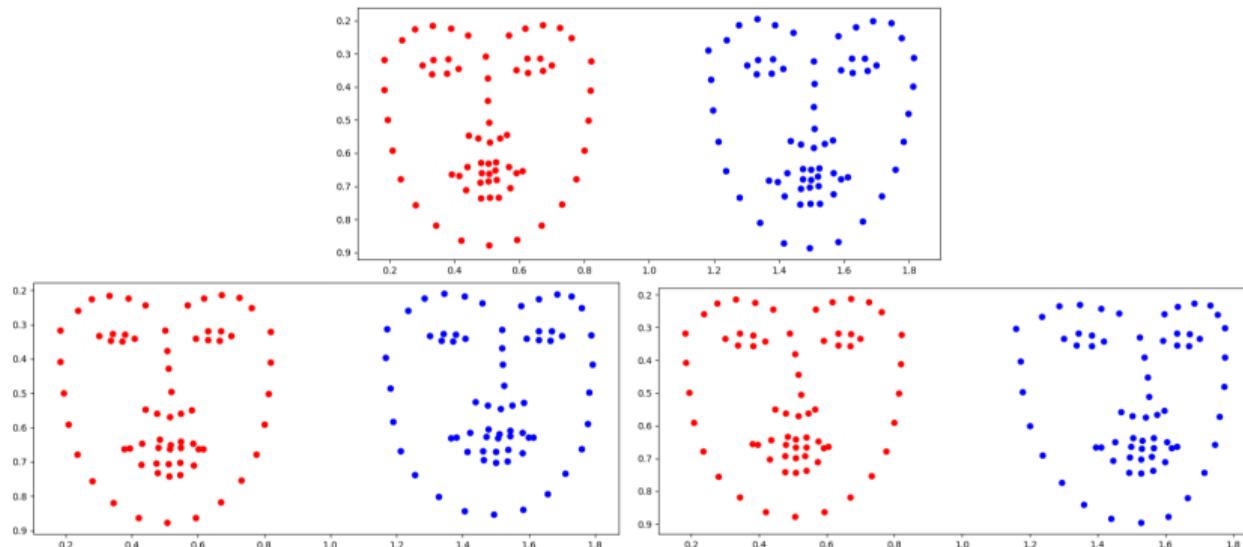
Với cách chuẩn hóa nêu trên, ta có thể đưa chuyển động của cột mốc gương mặt ban đầu lên cột mốc gương mặt chuẩn, với các góc độ quay khác nhau của video và loại bỏ đặc điểm gương mặt người nói



Hình 14: Kết quả chuẩn hóa cột mốc gương mặt. Cột mốc ban đầu (xanh), chuẩn hóa (đỏ)

Tiền xử lý dữ liệu

Nhờ việc điều chỉnh lại cột mốc gương mặt ở vùng mũi và miệng, vùng miệng của gương mặt được đưa về gần nhất với vùng miệng của cột mốc chuẩn, vì vậy tất cả các cột mốc được tính toán sẽ có cùng một kích thước vùng miệng, giúp cho mạng dễ học hơn



Hình 15: Kết quả chuẩn hóa vùng miệng. Cột mốc chưa chuẩn hóa (xanh), chuẩn hóa (đỏ)

Tiền xử lý dữ liệu

Dựa vào cột mốc gương mặt được trích xuất và chuẩn hóa, ta cũng sử dụng phép biến đổi Affine trên các khung hình trong video, để đưa gương mặt trên hình về vị trí khớp nhất với cột mốc chuẩn



Hình 16: Kết quả chuẩn hóa hình ảnh

Các tập dữ liệu được sử dụng

Tập dữ liệu GRID: Tập dữ liệu chứa 1000 câu, được nói bởi 34 người khác nhau. Như vậy, tập dữ liệu này chứa 34000 video với chất lượng cao, mỗi video có độ dài 3 giây.



Hình 17: Ảnh trích xuất từ các video trong tập dữ liệu GRID

Các tập dữ liệu được sử dụng

Tập dữ liệu LRW: Tập dữ liệu chứa hơn 1 triệu video khác nhau, mỗi video có độ dài 1.16 giây, gồm 29 khung hình. Những video này được chia làm 1000 từ vựng, mỗi từ vựng được nói bởi hơn 1000 người khác nhau. Tổng cộng, tập dữ liệu LRW chứa gần 1000 giờ video.



Hình 18: Ảnh trích xuất từ các video trong tập dữ liệu LRW

Các độ đo được sử dụng

Độ đo SSIM: Là độ đo sự tương đồng về mặt cảm quan giữa hai hình ảnh, dựa trên ba yếu tố là độ tương phản, ánh sáng, và cấu trúc của ảnh. Càng gần về 1, ảnh tạo sinh càng giống với ảnh gốc.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

Với:

- μ_x, μ_y : trung bình của các điểm ảnh trong cửa sổ
- σ_x^2, σ_y^2 : variance của các điểm ảnh trong cửa sổ
- C_1, C_2 : hằng số để ổn định phép chia

Các độ đo được sử dụng

Độ đo CPBD: Là độ đo thể hiện sự sắc nét của hình ảnh, là xác suất tích lũy của các cạnh trên ảnh nếu không được xem là cạnh mờ (blur). Một cạnh được xem là mờ khi nó có độ dày quá lớn, vượt qua độ dày được quy định là mờ W_{JNB}

$$CPBD = P(P_{BLUR} < P_{JNB}) = \sum_{P_{BLUR}=0}^{P_{BLUR}=P_{JNB}} (P_{BLUR}) \quad (6)$$

Với:

- P_{BLUR} : xác suất một cạnh bị mờ. $P_{BLUR} = 1 - e^{-(\frac{W_{e(i)}}{W_{JNB}(e_i)})^\beta}$
- W_{JNB} : Độ dày cạnh quy định là mờ
 - ▶ $W_{JNB} = 5$ nếu độ tương phản nhỏ hơn hoặc bằng 50
 - ▶ $W_{JNB} = 3$ nếu độ tương phản lớn hơn 50

Môi trường thí nghiệm

	Máy A	Máy B
CPU	Intel Xeon	Intel Core I5 8600K
Bộ nhớ	24GB	32GB
GPU	NVIDIA TESLA V100	NVIDIA Geforce RTX 3070
Ổ cứng	160GB SSD	768GB NVME SSD + 2TB HDD
OS	N/A(Colab Pro)	Ubuntu 20.04
Ngôn ngữ	Python 3.7	Python 3.7
Framework	Pytorch 1.8	Pytorch 1.8

Bảng 1: Các môi trường được sử dụng trong việc tiền xử lý dữ liệu, huấn luyện và thực hiện thí nghiệm

Thực hiện thí nghiệm

Quá trình huấn luyện mạng tạo sinh cột mốc gương mặt:

	GRID	LRW
Bộ tối ưu	Adam	Adam
Kích thước bó	100	100
Hệ số học	0.001	0.0002
Thời gian huấn luyện	5 phút	70 phút
$\mathcal{L}_{landmark}$	4.9200×10^{-4}	2.012×10^{-4}

Bảng 2: Chi tiết huấn luyện mạng tạo sinh cột mốc gương mặt. Giá trị mất mát (trên tập kiểm thử) và thời gian huấn luyện được ghi nhận tại vòng lặp cho ra mô hình tối ưu

Thực hiện thí nghiệm

Quá trình huấn luyện mạng GANs:

	GRID	LRW
Bộ tối ưu	Adam	Adam
Kích thước bộ	12	17
Hệ số học	0.0002	0.0002
Thời gian huấn luyện	5 giờ 20 phút	250 giờ
\mathcal{L}_{pix}	4.1575×10^{-2}	6.8502×10^{-2}
$\mathcal{L}_{gans-dis}$	0.9964	0.6931
$\mathcal{L}_{gans-landmark}$	7.6990×10^{-2}	5.2412×10^{-2}
\mathcal{L}	1.4892	1.4306

Bảng 3: Chi tiết huấn luyện mạng GANs. Giá trị măt măt (trên tập kiểm thử) và thời gian huấn luyện được ghi nhận tại vòng lặp cho ra mô hình tối ưu

1 Giới thiệu

2 Các công trình nghiên cứu có liên quan

- Mạng GANs
- Tổng quan tình hình nghiên cứu
- Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]

3 Phương pháp đề xuất

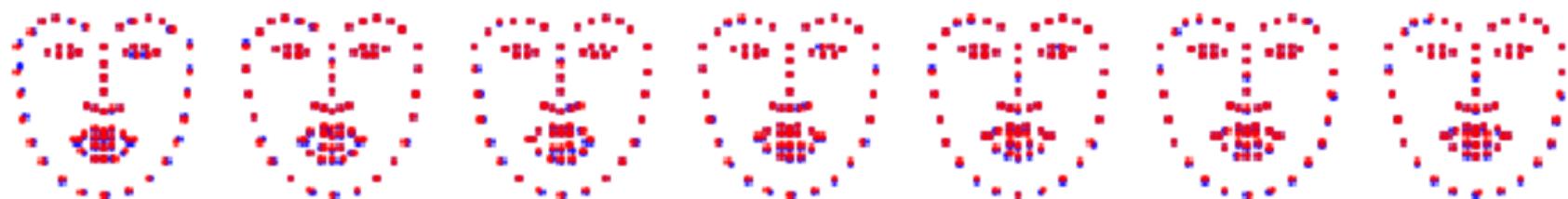
- Ý tưởng thực hiện luận văn
- Tiền xử lý dữ liệu
- Các tập dữ liệu được sử dụng
- Các độ đo được sử dụng
- Môi trường thí nghiệm
- Thực hiện thí nghiệm

4 Kết quả thí nghiệm

5 Kết luận

Kết quả thí nghiệm

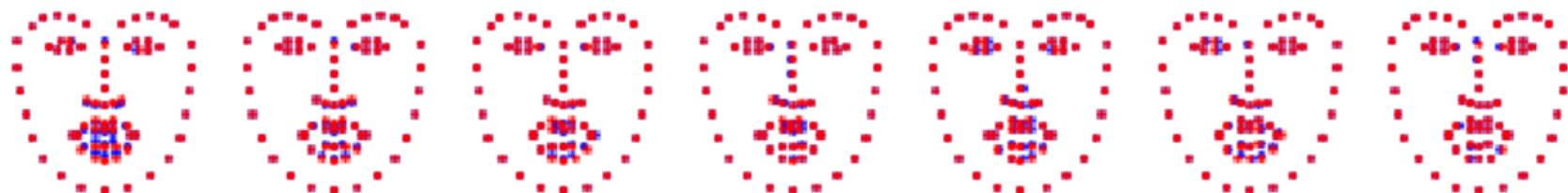
Tạo sinh cột mốc gương mặt trên tập GRID. Cột mốc tạo sinh (đỏ) bám tốt theo cột mốc gốc (xanh)



Hình 19: Kết quả tạo sinh cột mốc gương mặt theo giọng nói trên tập GRID. Cột mốc gốc màu xanh, cột mốc tạo sinh màu đỏ

Kết quả thí nghiệm

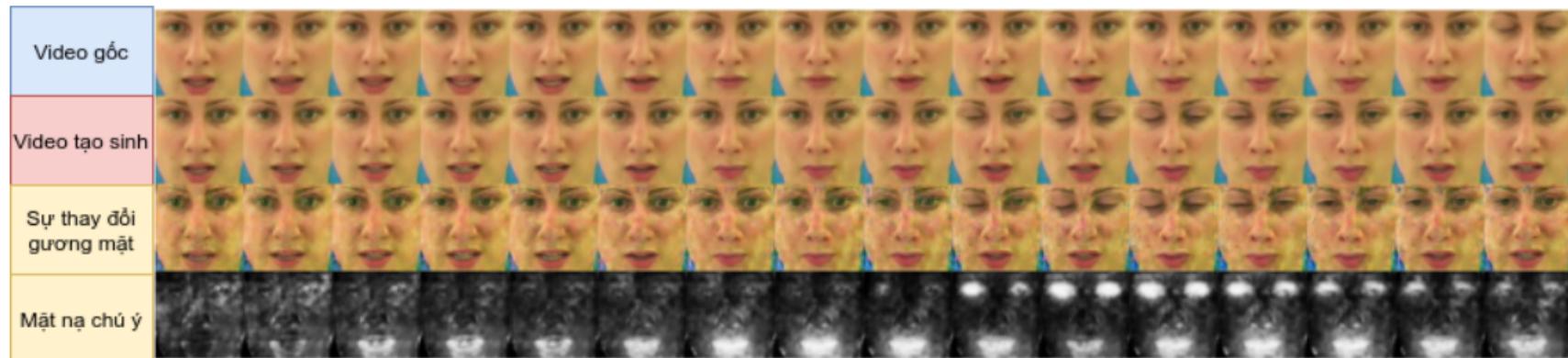
Tạo sinh cột mốc gương mặt trên tập LRW. Cột mốc tạo sinh (đỏ) bám tốt theo cột mốc gốc (xanh)



Hình 20: Kết quả tạo sinh cột mốc gương mặt theo giọng nói trên tập LRW. Cột mốc gốc màu xanh, cột mốc tạo sinh màu đỏ

Kết quả thí nghiệm

Tạo sinh gương mặt trên tập GRID. Hầu hết video được tạo sinh trên tập test đều rõ ràng, có chuyển động miệng hợp lý



Hình 21: Kết quả tạo sinh gương mặt theo giọng nói trên tập GRID

Kết quả thí nghiệm

Tạo sinh gương mặt trên tập LRW. Hầu hết video được tạo sinh tốt khi hình ảnh đầu vào là hình chụp thẳng mặt. Chất lượng video bị suy giảm khi hình ảnh đầu vào được chụp ở các góc nghiêng



Hình 22: Kết quả tạo sinh gương mặt theo giọng nói trên tập LRW

Kết quả thí nghiệm

Phương pháp	GRID		LRW	
	SSIM	CPBD	SSIM	CPBD
Zakharov et al [3]	0.54	0.19	0.42	0.11
Chung et al [4]	0.41	0.22	0.34	0.21
Baseline (Chen et al) [1]	0.41	0.08	0.38	0.07
Ours	0.72	0.12	0.54	0.06

Bảng 4: So sánh với các mạng có cùng mục tiêu về độ đo SSIM và CPBD. Dữ liệu trong bảng được lấy từ bài khảo sát [5]

Qua bảng trên ta thấy:

- Độ tương đồng SSIM đã tăng đáng kể cho cả hai tập dữ liệu. Do đó, video tạo sinh có chuyển động giống video gốc hơn
- Độ nét tăng đối với tập GRID, tuy nhiên lại giảm trên tập LRW

So sánh với mô hình gốc của Lele Chen



Hình 23: So sánh với mô hình của tác giả trên tập GRID. Bên trái là video được tạo sinh bởi mô hình mới, bên phải là của tác giả Lele Chen

So sánh với mô hình gốc của Lele Chen



Hình 24: So sánh với mô hình của tác giả trên tập LRW. Bên trái là video được tạo sinh bởi mô hình mới, bên phải là của tác giả Lele Chen

So sánh với mô hình gốc của Lele Chen

Về mặt cảm quan, ta thấy

- Đối với thử nghiệm trên tập GRID, hình ảnh được tạo sinh bởi mô hình mới có độ nét cao hơn
- Đối với thử nghiệm trên tập LRW, hình ảnh được tạo sinh bởi mô hình mới có độ nét thấp hơn
- Trên video thực tế, thử nghiệm trên cả hai tập cho kết quả tạo sinh mượt mà hơn và khẩu hình miệng khớp hơn trên mô hình mới

Điều này phù hợp với kết quả khảo sát về độ đo SSIM và CPBD

1 Giới thiệu

2 Các công trình nghiên cứu có liên quan

- Mạng GANs
- Tổng quan tình hình nghiên cứu
- Nghiên cứu Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss [1]

3 Phương pháp đề xuất

- Ý tưởng thực hiện luận văn
- Tiền xử lý dữ liệu
- Các tập dữ liệu được sử dụng
- Các độ đo được sử dụng
- Môi trường thí nghiệm
- Thực hiện thí nghiệm

4 Kết quả thí nghiệm

5 Kết luận

Kết luận

- Nghiên cứu đã chạy thử và kiểm chứng lại cấu trúc mạng của Lele Chen và cộng sự [1]
- Áp dụng có chỉnh sửa phương pháp chuẩn hóa cột mốc gương mặt từ bài báo [2] để cho ra kết quả tạo sinh hình ảnh tốt hơn so với bài báo gốc
- So sánh với các nghiên cứu mới, kết quả tạo sinh ảnh của mô hình vẫn còn kém hơn
- Tuy nhiên các nghiên cứu có kết quả tốt đòi hỏi quá nhiều tài nguyên tính toán, và khó có thể hiện thực trong điều kiện của tác giả
- Nghiên cứu có tiềm năng phát triển thêm để tạo thành một phần mềm tạo sinh biên tập viên ảo. Tác giả cũng đã tạo sinh thử hình ảnh của người Việt và âm thanh tiếng Việt
- Có thể phát triển thêm theo hướng sử dụng cột mốc gương mặt trong không gian 3 chiều, và tính toán chuyển động của đầu để video chân thực hơn

- [1] 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.
- [2] Sefik Emre Eskimez, Ross K. Maddox, Chenliang Xu, and Zhiyao Duan. Generating talking face landmarks from speech. *CoRR*, abs/1803.09803, 2018.
- [3] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *CoRR*, abs/1905.08233, 2019.
- [4] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *CoRR*, abs/1705.02966, 2017.
- [5] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark. *CoRR*, abs/2005.03201, 2020.

Thank you for your attention