Report paper: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

Giáp Ngọc Hiệu

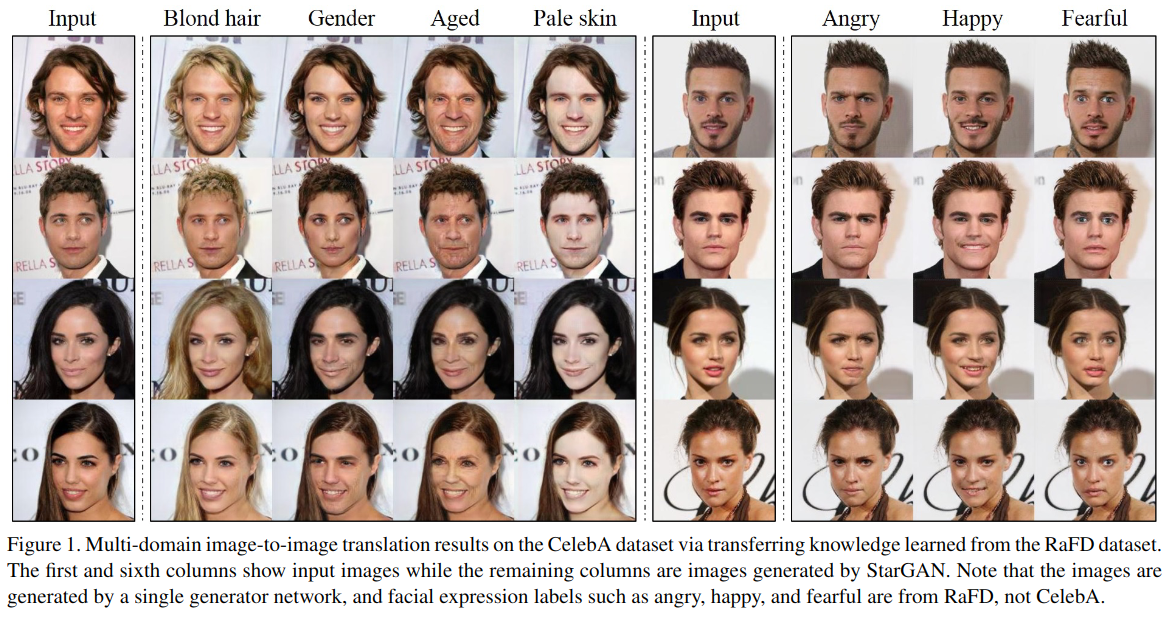
**Tóm tắt nội dung**

Bài report này là phần tóm tắt lại ý tưởng, nội dung chính, những điều làm được của paper

StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

1. **Ý tưởng**

Các nghiên cứu gần đây đã cho thấy sự thành công đáng kinh ngạc trong việc dịch từ hình ảnh sang hình ảnh cho hai lĩnh vực. Tuy nhiên, các phương pháp hiện có có khả năng mở rộng hạn chế và độ tin cậy hạn chế trong việc xử lý hơn hai lĩnh vực, vì các mô hình khác nhau phải được xây dựng độc lập cho mỗi cặp lĩnh vực hình ảnh. Để giải quyết hạn chế này, tác giả đề xuất StarGAN, một phương pháp mới và có khả năng mở rộng có thể thực hiện dịch từ hình ảnh sang hình ảnh cho nhiều lĩnh vực bằng chỉ một mô hình duy nhất. Kiến trúc mô hình thống nhất như vậy của StarGAN cho phép đào tạo đồng thời của nhiều bộ dữ liệu với các lĩnh vực khác nhau trong một mạng duy nhất. Điều này dẫn đến chất lượng ảnh được dịch của StarGAN vượt trội so với các mô hình hiện có cũng như khả năng mới mẻ của việc dịch linh hoạt một hình ảnh đầu vào sang bất kỳ lĩnh vực mục tiêu mong muốn nào.

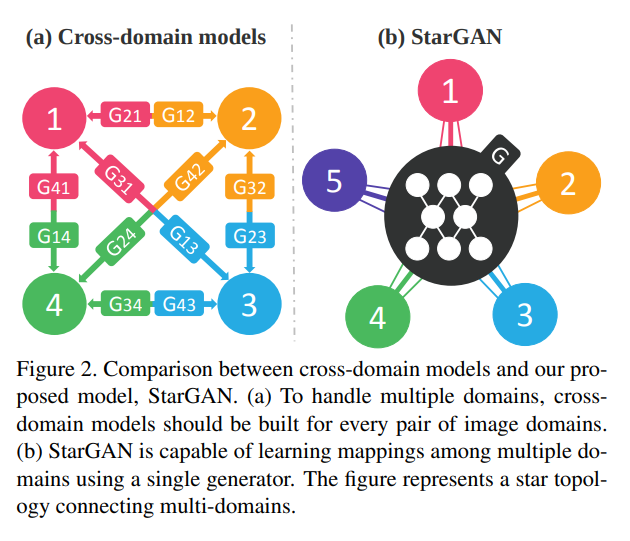


1. **Tổng quan**

Nhiệm vụ của việc dịch từ hình ảnh sang hình ảnh là thay đổi một khía cạnh cụ thể của một hình ảnh đã cho thành một khía cạnh khác, như thay đổi biểu cảm khuôn mặt từ cười thành cau có. Sự phát triển đáng kể đã xảy ra trong nhiệm vụ này sau khi giới thiệu các mạng GAN. Ví dụ, các nghiên cứu đã thể hiện khả năng thay đổi màu tóc, tái tạo hình ảnh từ bản đồ đường biên, và thay đổi mùa của hình ảnh phong cảnh. Đối với dữ liệu đào tạo từ hai lĩnh vực khác nhau, các mô hình học dịch các hình ảnh từ một lĩnh vực sang lĩnh vực khác. Thuật ngữ "thuộc tính" được đặt là một đặc điểm có ý nghĩa tồn tại trong một hình ảnh như màu tóc, giới tính hoặc tuổi, và "giá trị thuộc tính" là một giá trị cụ thể của một thuộc tính, ví dụ, đen/nâu cho màu tóc hoặc nam/nữ cho giới tính. Thuật ngữ "domain" được sử dụng để chỉ một tập hợp các hình ảnh chia sẻ cùng một giá trị thuộc tính, ví dụ, hình ảnh của phụ nữ có thể đại diện cho một domain trong khi những hình ảnh của đàn ông đại diện cho một domain khác.

Một số bộ dữ liệu hình ảnh đi kèm với các nhãn thuộc tính, ví dụ như bộ dữ liệu CelebA chứa 40 nhãn liên quan đến các đặc điểm khuôn mặt như màu tóc, giới tính và tuổi, và bộ dữ liệu RaFD có 8 nhãn cho biểu cảm khuôn mặt như 'vui vẻ', 'tức giận' và 'buồn bã'. Các cài đặt này cho phép thực hiện các nhiệm vụ thú vị hơn, như dịch từ hình ảnh sang hình ảnh đa miền, nơi có thể thay đổi hình ảnh theo các thuộc tính từ nhiều lĩnh vực khác nhau.

Tuy nhiên, các mô hình hiện có không hiệu quả và không hiệu quả trong các nhiệm vụ dịch hình ảnh đa miền như vậy. Sự không hiệu quả của chúng phát sinh từ việc cần phải học tất cả các ánh xạ giữa k lĩnh vực, đồng thời đào tạo k(k−1) bộ tạo ra riêng biệt. Mặc dù tồn tại các đặc điểm toàn cầu có thể được học từ hình ảnh của tất cả các lĩnh vực như hình dáng khuôn mặt, mỗi bộ tạo ra không thể sử dụng hoàn toàn toàn bộ dữ liệu đào tạo và chỉ có thể học từ hai lĩnh vực trong số k. Việc không sử dụng hoàn toàn dữ liệu đào tạo có thể giới hạn chất lượng của các hình ảnh được tạo ra. Hơn nữa, chúng không có khả năng đào tạo các lĩnh vực từ các bộ dữ liệu khác nhau cùng một lúc vì mỗi bộ dữ liệu được gán nhãn một cách một phần



Những đóng góp của tác giả bao gồm:

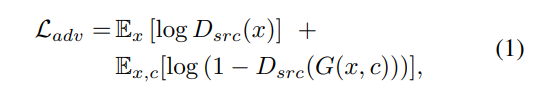
• Đề xuất StarGAN, một mạng mới có khả năng học cách ánh xạ giữa nhiều miền chỉ bằng một trình tạo duy nhất và một bộ phân biệt, huấn luyện hiệu quả từ hình ảnh của tất cả các miền.

• Chứng minh cách có thể học thành công việc dịch hình ảnh đa miền giữa nhiều bộ dữ liệu bằng cách sử dụng phương pháp mask vector cho phép StarGAN kiểm soát tất cả các nhãn miền có sẵn.

• Cung cấp cả kết quả định tính và định lượng về các tác vụ chuyển thuộc tính khuôn mặt và tổng hợp biểu hiện khuôn mặt bằng StarGAN, cho thấy tính ưu việt của nó so với các mô hình cơ sở.

1. **Phương pháp**
   1. **Multi-Domain Image-to-Image Translation**

* Mục tiêu là đào tạo một trình tạo G duy nhất học cách ánh xạ giữa nhiều miền. Để đạt được điều này, chúng ta huấn luyện G chuyển đổi hình ảnh đầu vào x thành hình ảnh đầu ra y theo điều kiện trên nhãn miền đích c, G(x, c) → y. Chúng ta tạo ngẫu nhiên nhãn miền đích c để G học cách dịch hình ảnh đầu vào một cách linh hoạt. Ngoài ra còn có một bộ phân loại cho phép một bộ phân biệt duy nhất kiểm soát nhiều miền. Nghĩa là, trình phân biệt tạo ra các phân bố xác suất trên cả nguồn và nhãn miền, D : x → {Dsrc(x), Dcls(x)}
* **Adversarial Loss.**

****

trong đó G tạo ra một ảnh G(x, c) dựa trên cả ảnh đầu vào x và nhãn miền đích c, trong khi D cố gắng phân biệt giữa ảnh thật và ảnh giả. Trong bài báo này, Dsrc(x) là phân bố xác suất trên các nguồn do D đưa ra. G cố gắng giảm thiểu mục tiêu này, trong khi D cố gắng tối đa hóa nó.

* **Domain Classification Loss.**

Đối với một hình ảnh đầu vào x nhất định và một nhãn miền đích c, mục tiêu của chúng ta là chuyển x thành một hình ảnh đầu ra y, được phân loại chính xác cho miền đích c. Để đạt được điều kiện này, chúng ta thêm một trình phân loại phụ trợ lên trên D và áp đặt domain classification loss khi tối ưu hóa cả D và G. Nghĩa là, chúng ta phân tách mục tiêu thành: domain classification loss của hình ảnh thực được sử dụng để tối ưu hóa D, và domain classification loss của các hình ảnh giả được sử dụng để tối ưu hóa G.

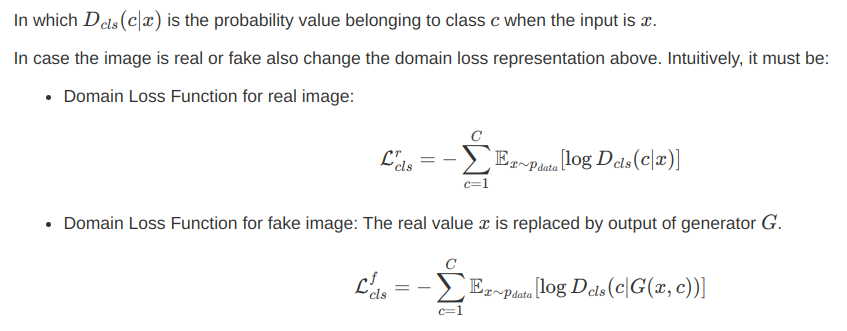
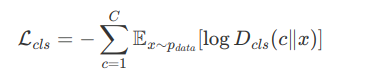


trong đó Dcls(c′|x) biểu thị phân bố xác suất trên các nhãn miền được tính toán bởi D. Bằng cách cực tiểu hóa mục tiêu này, D học cách phân loại một ảnh thực x vào miền gốc tương ứng c′ của nó. Chúng ta giả định rằng cặp nhãn tên miền và hình ảnh đầu vào (x, c′) được cung cấp bởi dữ liệu huấn luyện. Mặt khác, hàm mất mát để phân loại miền của hình ảnh giả được định nghĩa là



Nói cách khác, G cố gắng giảm thiểu mục tiêu này để tạo ra các ảnh có thể được phân loại là miền đích c.

Domain Classification Loss chỉ nhằm mục đích tối ưu hóa vấn đề Multi-Domain classification. Các nhãn mà bộ phân loại Multi-Domain cần dự đoán là các đặc điểm miền đích của ảnh giả hoặc các đặc điểm thật của ảnh thật. Nói chung, hàm này có dạng cross entropy loss function:



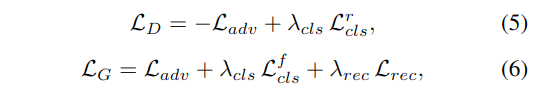
* **Reconstruction Loss.**

Bằng cách giảm thiểu adversarial và classification losses, G được đào tạo để tạo ra hình ảnh chân thực và được phân loại theo miền mục tiêu chính xác. Tuy nhiên, việc sử dụng 2 loss không đảm bảo rằng hình ảnh được dịch sẽ giữ nguyên nội dung của hình ảnh đầu vào trong khi chỉ thay đổi phần liên quan đến miền của đầu vào. Để giảm bớt vấn đề này, tác giả áp dụng cycle consistency loss cho G, được định nghĩa là:



trong đó G lấy ảnh đã dịch G(x, c) và nhãn miền gốc c′ làm đầu vào và cố gắng xây dựng lại ảnh gốc x. L1 norm được áp dụng cho loss này. Lưu ý rằng G được dùng hai lần, lần đầu tiên để dịch ảnh gốc thành ảnh trong miền đích và sau đó để tái tạo lại ảnh gốc từ ảnh đã dịch.

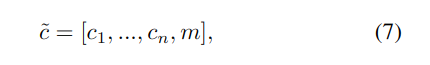
* **Full Objective.**



trong đó λcls và λrec là các siêu tham số kiểm soát tầm quan trọng tương đối của domain classification và reconstruction losses. Tác giả sử dụng λcls = 1 và λrec = 10 trong tất cả các thử nghiệm của mình.

* 1. **Training with Multiple Datasets**
* Ưu điểm quan trọng của StarGAN là nó kết hợp đồng thời nhiều bộ dữ liệu chứa các loại nhãn khác nhau để StarGAN có thể kiểm soát tất cả các nhãn ở giai đoạn test. Tuy nhiên, một vấn đề khi học từ nhiều tập dữ liệu là thông tin nhãn chỉ được biết một phần đối với mỗi tập dữ liệu. Trong trường hợp CelebA và RaFD, trong khi CelebA chứa nhãn cho các thuộc tính như màu tóc và giới tính, thì nó không có bất kỳ nhãn nào cho biểu cảm khuôn mặt như 'vui vẻ' và 'tức giận' và ngược lại của RaFD. Đây là vấn đề vì thông tin đầy đủ trên vectơ nhãn c′ là cần thiết khi xây dựng lại hình ảnh đầu vào x từ hình ảnh được dịch G(x, c)
* **Mask Vector.**

Để giảm bớt vấn đề này, tác giả giới thiệu vectơ mặt nạ m cho phép StarGAN bỏ qua các nhãn không xác định và tập trung vào nhãn được biết rõ ràng do một tập dữ liệu cụ thể cung cấp. Trong StarGAN, tác giả sử dụng n-dimensional one-hot vector để biểu thị m, với n là số lượng tập dữ liệu. Ngoài ra, tác giả xác định phiên bản hợp nhất của nhãn dưới dạng vectơ:



trong đó [·] đề cập đến phép nối và ci đại diện cho một vectơ cho các nhãn của tập dữ liệu thứ i. Vectơ của nhãn đã biết ci có thể được biểu diễn dưới dạng vectơ nhị phân cho các thuộc tính nhị phân hoặc vectơ one-hot cho các thuộc tính phân loại. Đối với n-1 nhãn chưa biết còn lại, chúng ta chỉ cần gán giá trị 0. Trong các thử nghiệm của mình, tác giả sử dụng bộ dữ liệu CelebA và RaFD, trong đó n là hai.

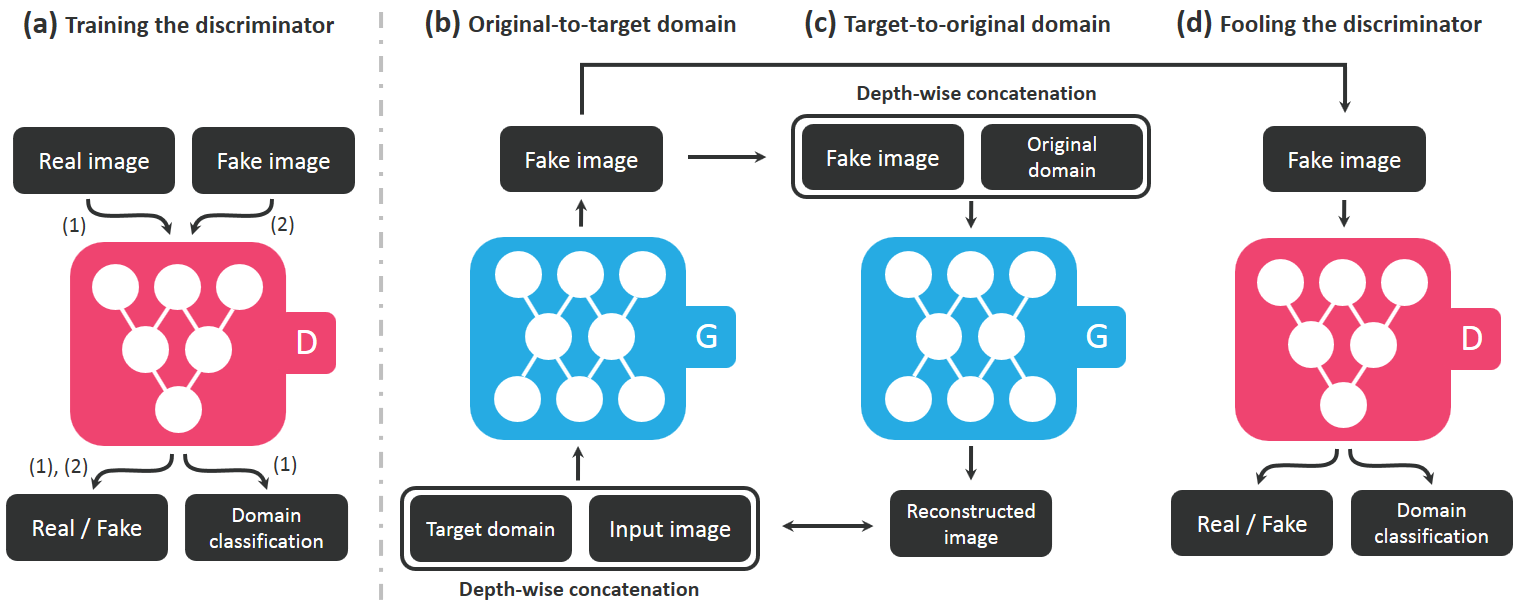
* **Training** **Strategy**.

Khi đào tạo StarGAN với nhiều tập dữ liệu, chúng ta sử dụng nhãn miền ̃c được xác định trong biểu thức (7) làm đầu vào cho G. Bằng cách đó, G học cách bỏ qua các nhãn không xác định, là các vectơ bằng 0 và tập trung vào nhãn được đưa ra rõ ràng. Cấu trúc của G hoàn toàn giống như khi huấn luyện với một tập dữ liệu duy nhất, ngoại trừ kích thước của nhãn đầu vào ̃c. Mặt khác, mở rộng bộ phân loại của D để tạo ra phân bố xác suất trên các nhãn cho tất cả các tập dữ liệu. Sau đó, huấn luyện mô hình trong cài đặt học tập đa tác vụ, trong đó D cố gắng chỉ giảm thiểu lỗi phân loại liên quan đến nhãn đã biết. Ví dụ: khi đào tạo với hình ảnh trong CelebA, D chỉ giảm thiểu các lỗi phân loại cho các nhãn liên quan đến thuộc tính CelebA chứ không phải biểu cảm khuôn mặt liên quan đến RaFD. Trong các cài đặt này, bằng cách luân phiên giữa CelebA và RaFD, D sẽ tìm hiểu tất cả các đặc trưng cho cả hai tập dữ liệu và G sẽ học cách kiểm soát tất cả các nhãn trong cả hai tập dữ liệu.

* 1. **Flow**
* **Training within a Single Dataset**

Tổng quan về StarGAN bao gồm một bộ phân biệt D và một bộ tạo G:

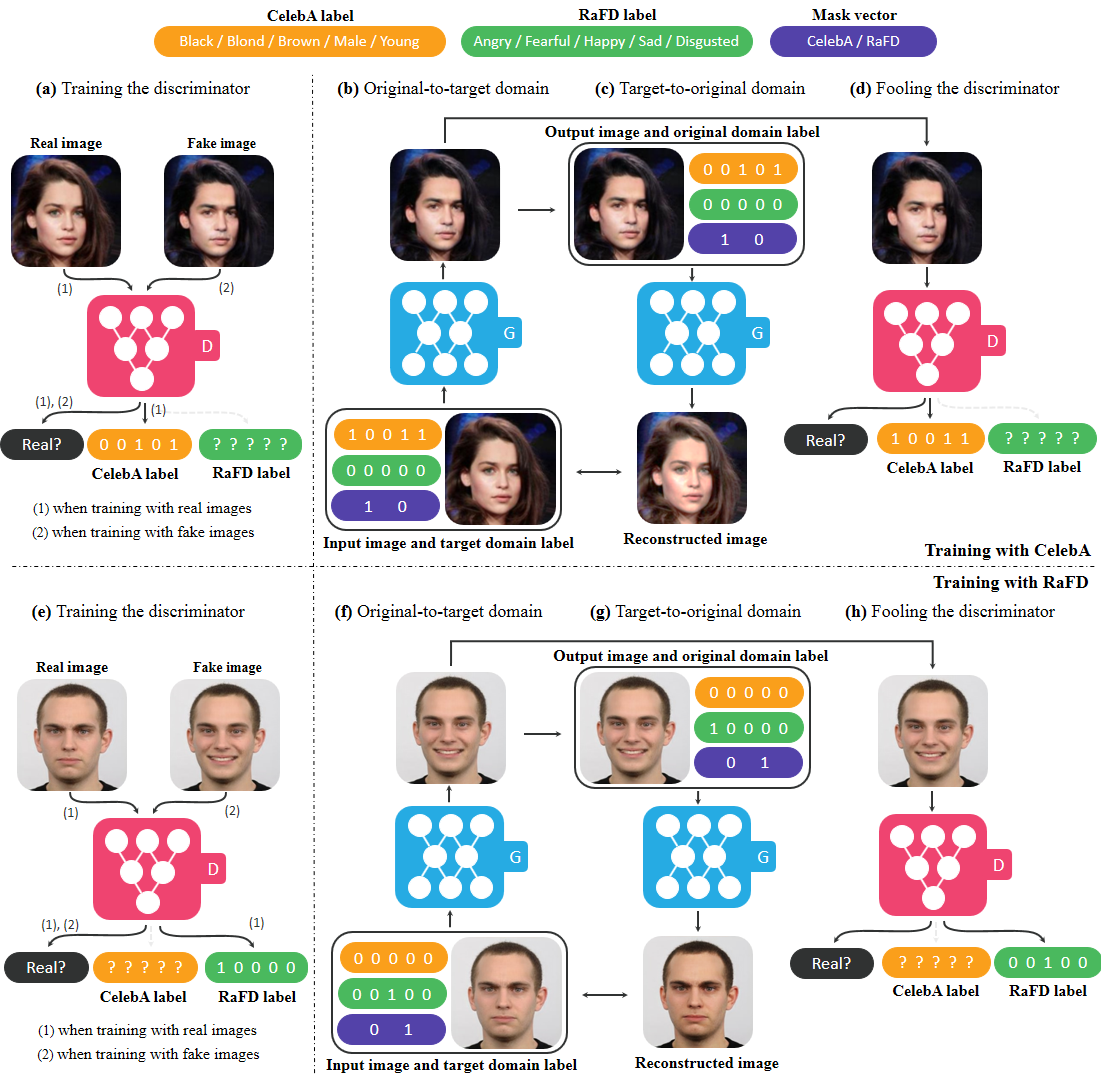
* (a) D học để phân biệt giữa hình ảnh thật và giả và phân loại các hình ảnh thật vào miền tương ứng của chúng.
* (b) G nhận hình ảnh và nhãn miền mục tiêu làm đầu vào và tạo ra một hình ảnh giả. Nhãn miền mục tiêu được sao chép không gian và nối với hình ảnh đầu vào.
* (c) G cố gắng tái tạo hình ảnh gốc từ hình ảnh giả với nhãn miền ban đầu.
* (d) G cố gắng tạo ra hình ảnh không thể phân biệt được từ hình ảnh thật và có thể được D phân loại vào miền mục tiêu.

****

* **Training with Multiple Datasets**

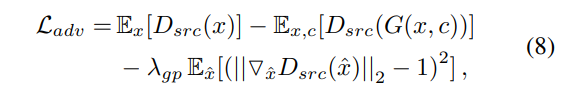
Tổng quan về StarGAN khi đào tạo với cả CelebA và RaFD:

* (a) ~ (d) mô tả quá trình đào tạo sử dụng CelebA, và (e) ~ (h) mô tả quá trình đào tạo sử dụng RaFD.
* (a), (e): D học để phân biệt giữa hình ảnh thật và giả và giảm thiểu lỗi phân loại chỉ cho nhãn đã biết.
* (b), (c), (f), (g): Khi mask vector (màu tím) là [1, 0], G học tập trung vào nhãn CelebA (màu vàng) và bỏ qua nhãn RaFD (màu xanh lá cây) để thực hiện dịch hình ảnh sang hình ảnh, và ngược lại khi mask vector là [0, 1].
* (d), (h): G cố gắng tạo ra hình ảnh không thể phân biệt được với hình ảnh thật và có thể được D phân loại vào miền mục tiêu.



* 1. **Implementation**
* **Improved GAN Training.**

Để ổn định quá trình đào tạo và tạo ra hình ảnh chất lượng cao hơn, tác giả thay thế phương trình (1) với Wasserstein GAN objective with gradient penalty được xác định là

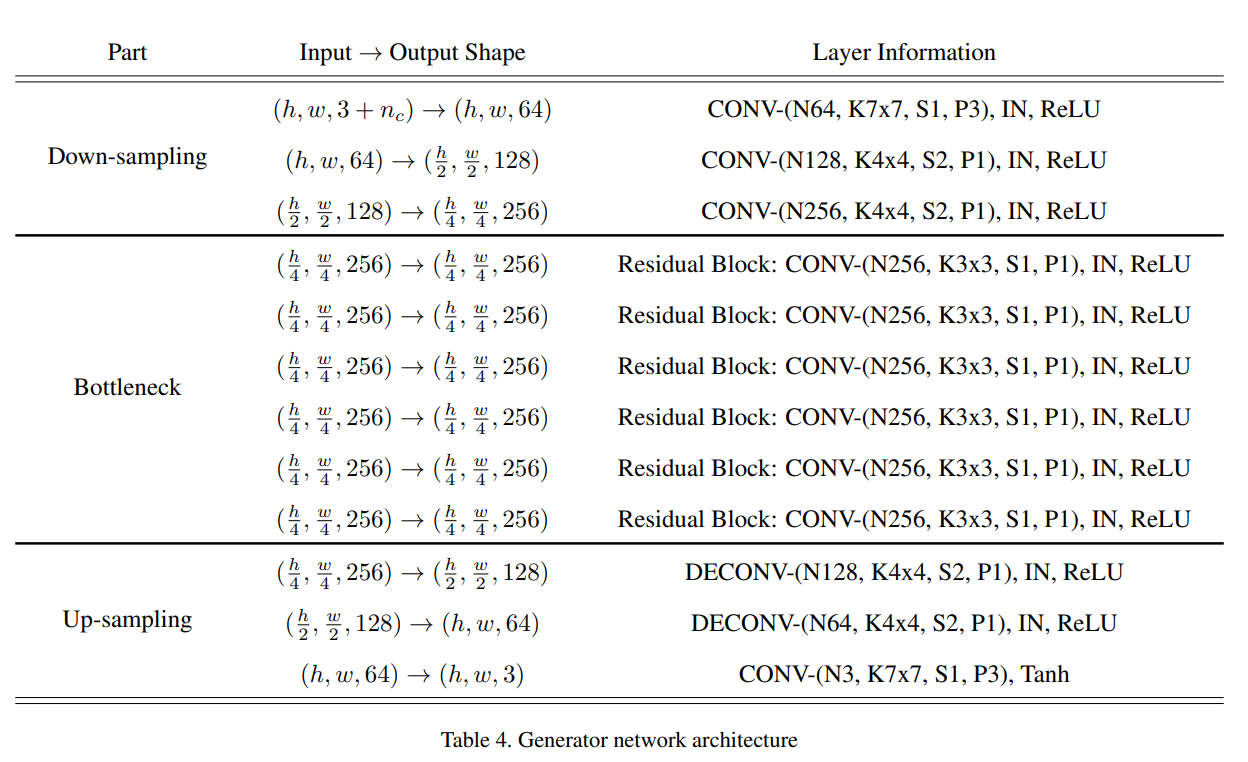


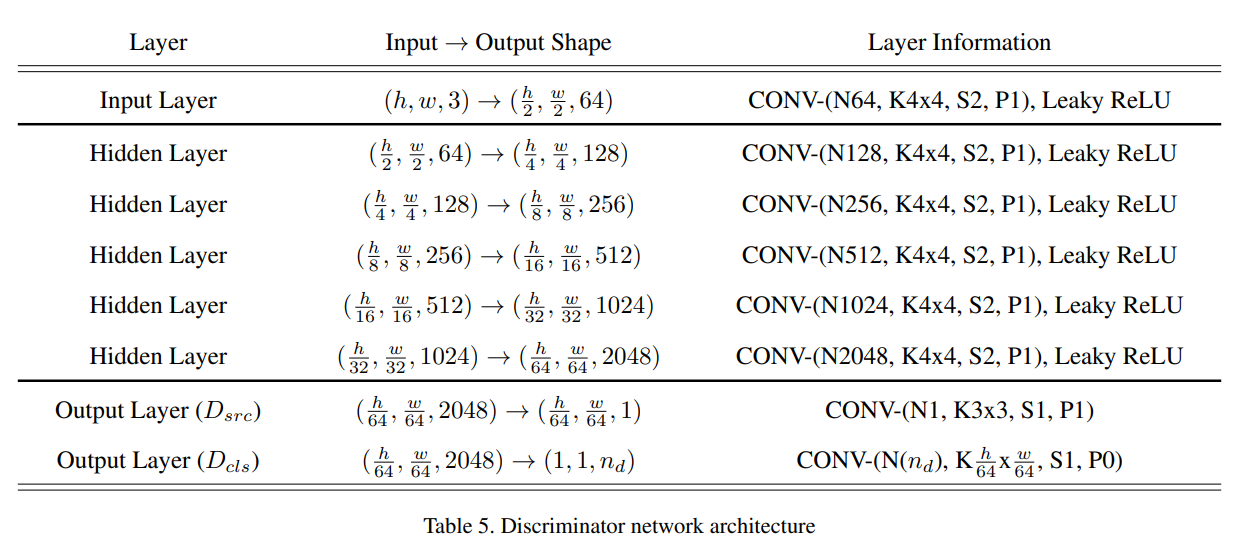
trong đó ˆx được lấy mẫu thống nhất dọc theo một đường thẳng giữa một cặp ảnh thật và ảnh được tạo. Tác giả sử dụng λgp = 10 cho tất cả các thử nghiệm.

* **Network Architecture.**

Được điều chỉnh từ CycleGAN, StarGAN có G bao gồm 2 convolutional layers với stride=2 cho downsampling, 6 residual blocks và 2 transposed convolutional layers với stride=2 cho upsampling. Tác giả sử dụng instance normalization cho G nhưng không sử dụng cho D. Áp dụng PatchGAN cho D, phân loại xem các bản vá hình ảnh cục bộ là thật hay giả.

Dây là một số ký hiệu; nd: số lượng miền, nc: kích thước của nhãn miền (nd + 2 khi đào tạo với cả bộ dữ liệu CelebA và RaFD, nếu không thì giống như nd), N: số lượng kênh đầu ra, K: kernel size, S: stride size, P: padding size, IN: instance normalization





1. **Results**
   1. **Dataset**

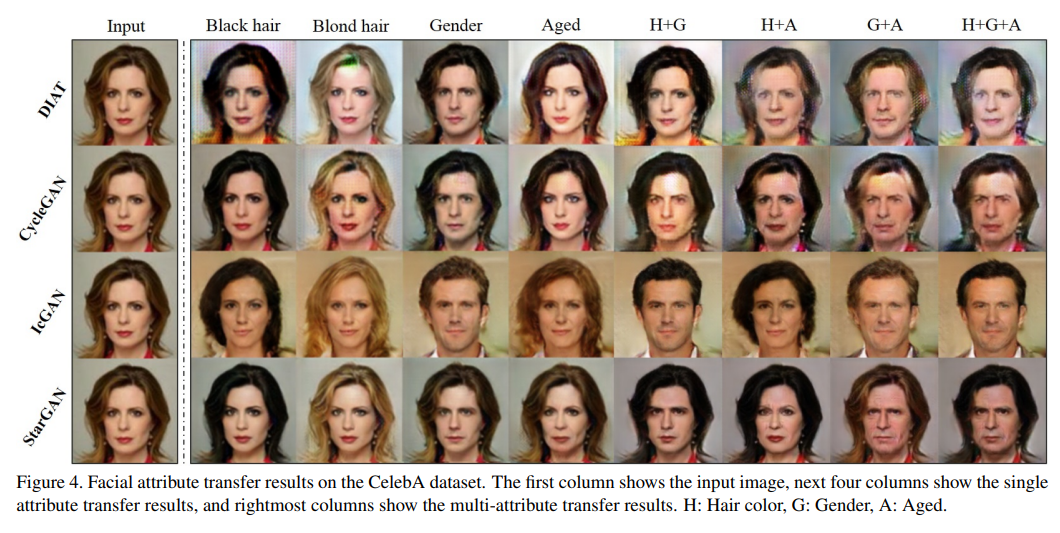
* **CelebA**.

Bộ dữ liệu CelebA) chứa 202.599 hình ảnh khuôn mặt của những người nổi tiếng, mỗi hình ảnh được chú thích bằng 40 thuộc tính nhị phân. Trong paper, tác giả cắt các hình ảnh có kích thước 178 × 218 ban đầu thành 178 × 178, sau đó thay đổi kích thước chúng thành 128 × 128. Chọn ngẫu nhiên 2.000 hình ảnh làm tập kiểm tra và sử dụng tất cả các hình ảnh còn lại cho dữ liệu huấn luyện. Tác giả xây dựng bảy miền sử dụng các thuộc tính sau: màu tóc (đen, vàng, nâu), giới tính (nam/nữ) và tuổi (trẻ/già)

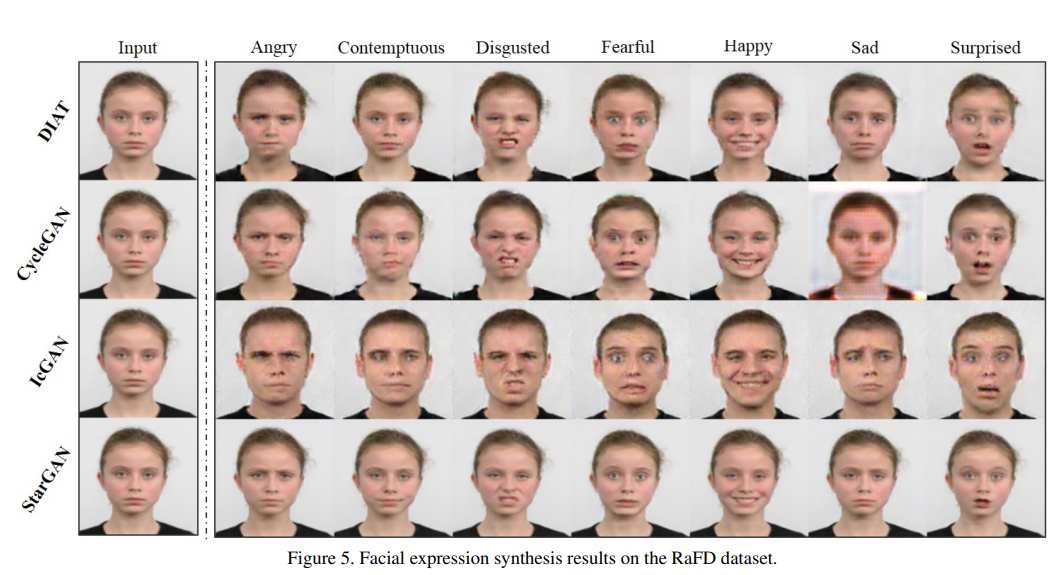
* **RaFD**

RaFD bao gồm 4.824 hình ảnh được thu thập từ 67 người tham gia. Mỗi người tham gia thực hiện tám nét mặt theo ba hướng nhìn khác nhau, được ghi lại từ ba góc độ khác nhau. Tác giả cắt hình ảnh thành 256x256, trong đó các khuôn mặt được căn giữa và sau đó thay đổi kích thước chúng thành 128x128.

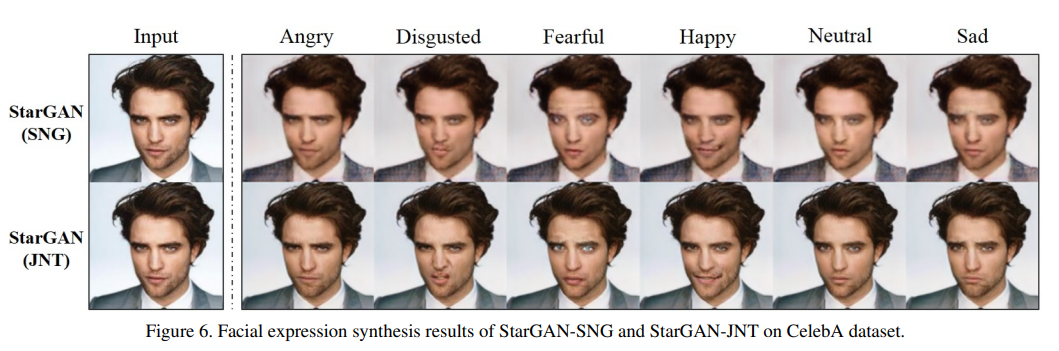
* 1. **Some Training Details**
* Tất cả các mô hình đều được huấn luyện bằng Adam với β1 = 0,5 và β2 = 0,999. Để tăng cường dữ liệu, tác giả lật hình ảnh theo chiều ngang với xác suất là 0,5. Tác giả thực hiện một lần cập nhật trình tạo sau năm lần cập nhật bộ phân biệt. Kích thước lô được đặt thành 16 cho tất cả các thử nghiệm. Đối với các thử nghiệm trên CelebA, tác giả đào tạo tất cả các mô hình với tỷ lệ học tập là 0,0001 trong 10 kỷ nguyên đầu tiên và giảm tuyến tính tỷ lệ học tập xuống 0 trong 10 kỷ nguyên tiếp theo. Để bù đắp cho việc thiếu dữ liệu, khi đào tạo bằng RaFD, tác giả đào tạo tất cả các mô hình trong 100 kỷ nguyên với tốc độ học tập là 0,0001 và áp dụng chiến lược phân rã tương tự trong 100 kỷ nguyên tiếp theo. Quá trình đào tạo mất khoảng một ngày trên một GPU NVIDIA Tesla M40.
  1. **Results**
* **Qualitative Results**

****

StarGAN cung cấp chất lượng hình ảnh dịch cao hơn trên dữ liệu kiểm tra so với các cross-domain models.



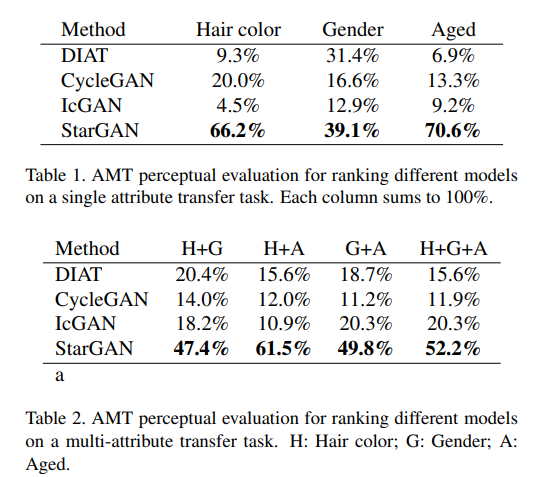
StarGAN rõ ràng tạo ra các biểu hiện tự nhiên nhất trong khi duy trì đúng bản chất cá nhân và đặc điểm khuôn mặt của hình ảnh đầu vào.



Để phân biệt giữa mô hình chỉ được đào tạo trên RaFD và mô hình được đào tạo trên cả CelebA và RaFD, mô hình trước được ký hiệu là StarGAN-SNG (đơn) và mô hình sau được ký hiệu là StarGAN-JNT (chung).

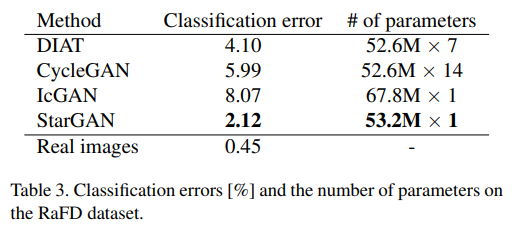
Bằng cách sử dụng cả CelebA và RaFD, StarGAN-JNT có thể cải thiện các tác vụ cấp thấp này, điều này có lợi cho việc học tổng hợp biểu cảm khuôn mặt.

* **Quantitative Results**

****

Amazon Mechanical Turk (AMT) được sử dụng để đánh giá single and multiple attribute transfer tasks.

StarGAN nhận được đa số phiếu bầu cho thuộc tính chuyển giao tốt nhất trong mọi trường hợp.



StarGAN nhỏ hơn DIAT 7 lần và nhỏ hơn CycleGAN 14 lần.

1. **Note**

* **Ưu điểm**:
  + Tính linh hoạt : StarGAN mang đến sự linh hoạt bằng cách cho phép thao tác đồng thời nhiều thuộc tính trong một hình ảnh. Điều này rất hữu ích trong các ứng dụng khác nhau như chỉnh sửa thuộc tính khuôn mặt, thay đổi kiểu tóc hoặc thậm chí tạo ra hình ảnh với các đặc điểm đa dạng.
  + Unified Model : Việc sử dụng mô hình hợp nhất giúp đơn giản hóa quá trình huấn luyện và giảm số lượng tham số so với các mô hình yêu cầu mạng riêng cho từng miền. Điều này có thể tiết kiệm tài nguyên tính toán và thời gian đào tạo.
  + Chất lượng và tính chân thực : Sự kết hợp giữa adversarial loss and cycle consistency loss giúp tạo ra hình ảnh chân thực và chất lượng cao trong quá trình dịch tên miền.
* **Nhược điểm**:
  + Độ phức tạp : Mặc dù StarGAN đơn giản hóa quá trình đào tạo so với một số lựa chọn thay thế, nhưng đây vẫn là một mô hình tương đối phức tạp và việc hiểu được sự phức tạp của nó có thể là một thách thức.
  + Thử thách đào tạo : Việc đào tạo GAN có thể là một thử thách và StarGAN cũng không ngoại lệ.
  + Yêu cầu dữ liệu : Giống như các mô hình học sâu khác, StarGAN yêu cầu một lượng đáng kể dữ liệu được gắn nhãn cho các miền và thuộc tính khác nhau để đạt được kết quả khả quan.
  + Artefact Generation : Trong một số trường hợp, StarGAN có thể tạo ra đồ giả hoặc đặc điểm không thực tế trong hình ảnh được dịch. Đây có thể là vấn đề thường gặp trong các tác vụ dịch từ hình ảnh sang hình ảnh bằng GAN.

-------------------Ending Mệt vãi ---------------------