Nguyễn Danh Nghi 15110087

Huỳnh Duy Anh 15110004

One-Sided Unsupervised Domain Mapping

# **TÓM TẮT**

Trong ánh xạ miền không giám sát, người học được cung cấp hai tập dữ liệu chưa có A và B. Mục tiêu là tìm hiểu một ánh xạ GAB dịch một mẫu trong A sang mẫu tương tự trong B. Các phương pháp gần đây đã chỉ ra rằng khi học đồng thời cả GAB và nghịch đảo lập ánh xạ GBA, ánh xạ thuyết phục thu được. Trong công việc này, chúng tôi trình bày một phương pháp học GAB mà không cần học GBA. Điều này được thực hiện bằng cách học một ánh xạ duy trì khoảng cách giữa một cặp mẫu. Hơn nữa, ánh xạ tốt thu được, thậm chí bằng cách duy trì khoảng cách giữa các phần khác nhau của cùng một mẫu trước và sau khi ánh xạ. Chúng tôi trình bày các kết quả thử nghiệm rằng phương pháp mới không chỉ cho phép học tập lập ánh xạ một mặt, mà còn dẫn đến các kết quả số thích hợp hơn so với ràng buộc dựa trên chu kỳ hiện tại. Toàn bộ mã của chúng tôi được cung cấp công khai tại   
<https://github.com/sagiebenaim/DistanceGAN>

# **GIỚI THIỆU**

Sự ra đời của công nghệ Generative Adversarial Network (GAN) đã cho phép tạo ra các hình ảnh thực tế bắt chước một tập huấn nhất định bằng cách nắm bắt chính xác những gì nằm trong lớp đã cho và những gì là “giả”. Trong số nhiều nhiệm vụ được thực hiện bởi GAN, nhiệm vụ ánh xạ một hình ảnh trong miền nguồn tới hình ảnh tương tự trong một miền đích là một lợi ích cụ thể.

Các giải pháp được đề xuất cho vấn đề này thường có thể được phân tách bằng số lượng giám sát bắt buộc. Trên cùng một phương pháp, các phương pháp được giám sát hoàn toàn sử dụng các cặp mẫu phù hợp, một trong mỗi miền, để tìm hiểu ánh xạ. Giám sát trực tiếp ít hơn đã được chứng minh bằng cách sử dụng một ánh xạ vào một không gian ngữ nghĩa và yêu cầu mẫu ban đầu và mẫu tương tự trong miền đích chia sẻ cùng một đại diện ngữ nghĩa.

Nếu hai tên miền có liên quan cao, nó đã được chứng minh rằng chỉ bằng cách chia sẻ trọng số giữa các mạng hoạt động trên hai miền và không cần giám sát thêm, người ta có thể ánh xạ mẫu giữa hai miền. Đối với các miền xa hơn, nó đã được chứng minh gần đây rằng bằng ánh xạ đối xứng nghiêng theo cả hai hướng, lấy vật có ý nghĩa tương tự. Điều này được thực hiện bằng cách yêu cầu tính tròn, tức là, ánh xạ mẫu từ miền này sang miền khác và sau đó quay lại, tạo mẫu gốc.

Trong công việc này, chúng ta tiến thêm một bước nữa và cho thấy rằng có thể tìm hiểu ánh xạ giữa miền nguồn và tên miền đích theo cách không giám sát một chiều, bằng cách thực thi tương quan chéo giữa các khoảng cách tương ứng được tính toán trong mỗi miền. Ràng buộc mới cho phép lập ánh xạ một chiều và cũng cung cấp, trong các thí nghiệm của chúng tôi, kết quả số tốt hơn so với tính tròn. Kết hợp cả hai ràng buộc này với nhau thường dẫn đến những cải tiến hơn nữa.

Việc tìm hiểu ràng buộc mới đòi hỏi phải so sánh các cặp mẫu. Mặc dù không có lý do thực tế nào để không làm như vậy, vì lẽ rằng các tập huấn luyện gồm có nhiều mẫu vật, chúng tôi chứng minh rằng các ràng buộc tương tự thậm chí có thể được áp dụng cho mỗi hình ảnh bằng cách tính toán khoảng cách giữa, ví dụ: phần trên cùng của hình ảnh và phần dưới cùng.

## **Related work (Công việc có liên quan)**

**Style transfer (Chuyển kiểu)**

Những phương thức này thường nhận được như là đầu vào một dạng hình ảnh và một nội dung hình ảnh và tạo ra một hình ảnh mới có kiểu đầu tiên và nội dung thứ hai. Vấn đề dịch hình ảnh giữa các tên miền khác nhau vì khi ánh xạ giữa các tên miền, một phần của nội dung được thay thế bằng nội dung mới khớp với tên miền đích và không chỉ là kiểu. Tuy nhiên, sự khác biệt là không rõ ràng, và nhiều ví dụ ánh xạ tên miền chéo (cross-domain) trong các tài liệu hầu như có thể được xem như là chuyển giao kiểu (Style Transfer). {Còn ví dụ chưa dịch xong}

**Generative Adversarial Networks (Có khả năng sinh ra những Mạng Đối Nghịch - GAN)**

Các phương thức GAN huấn luyện một mạng sinh G để tổng hợp các mẫu từ một phân phối mục tiêu, các vectơ nhiễu xác định, cùng nhau huấn luyện mạng thứ hai D. Kiến trúc sinh trưởng cụ thể mà chúng ta và người khác sử dụng dựa trên kiến trúc. Trong ánh xạ hình ảnh, hình ảnh được tạo dựa trên hình ảnh đầu vào chứ không dựa trên nhiễu ngẫu nhiên.

**Unsupervised Mapping (Ánh xạ không giám sát)**

Công việc liên quan nhiều nhất đến công việc của chúng tôi, không sử dụng giám sát ngoại trừ các hình ảnh mẫu từ hai miền. Điều này đã được thực hiện rất gần đây trong hình ảnh để dịch hình ảnh và hơi sớm hơn để dịch giữa các ngôn ngữ tự nhiên. Lưu ý rằng đề xuất phương pháp "GAN với khôi phục mất mác", áp dụng hạn chế chu kỳ ở một bên và chỉ đào tạo một GAN. Tuy nhiên, không giống như phương thức của chúng ta, phương thức này đòi hỏi sự phục hồi của cả hai ánh xạ và được thực hiện tốt hơn bằng phương thức hai chiều đầy đủ.

Phương thức CoGAN, tìm hiểu ánh xạ từ một vector đầu vào ngẫu nhiên đến các mẫu phù hợp từ hai miền. Nó đã được chứng minh rằng phương pháp có thể được sửa đổi để thực hiện dịch miền. Trong CoGAN, hai miền được giả định là tương tự nhau và bộ sinh của chúng (và các thiết bị so sánh GAN) chia sẻ nhiều trọng số lớp, tương tự như. Như đã được chứng minh trong, phương pháp không cạnh tranh trong lĩnh vực hình ảnh để dịch hình ảnh.

**Weakly Supervised Mapping (Ánh xạ yếu được giám sát)**

Sự kết hợp giữa miền nguồn và miền đích được thực hiện bằng cách kết hợp một bản đồ cố định được huấn luyện trước f và yêu cầu f-constancy, tức là các hoạt động của f giống nhau đối với các mẫu đầu vào và các mẫu được ánh xạ.

**Supervised Mapping (Ánh xạ được giám sát)**

Khi được cung cấp các cặp kết hợp (hình ảnh đầu vào, hình ảnh đầu ra), việc giám sát có thể được thực hiện trực tiếp. Một ví dụ về phương pháp đó cũng sử dụng GAN, là nơi mà người phân biệt D nhận được một cặp hình ảnh có một hình ảnh nguồn và hình ảnh kia là hình ảnh đích (cặp "thực") hoặc hình ảnh được tạo ra (cặp "giả"). Việc liên kết giữa nguồn và ảnh đích được tăng cường hơn nữa bằng cách sử dụng kiến trúc U-net.

**Domain adaptation (Sự thích ứng miền)**

Trong cài đặt này, chúng tôi thường được cung cấp hai tên miền, một miền có giám sát ở dạng nhãn phù hợp, trong khi thứ hai có ít hoặc không có giám sát. Mục tiêu là tìm hiểu cách gắn nhãn các mẫu từ tên miền thứ hai. những gì là phổ biến cho cả hai lĩnh vực và những gì là riêng biệt được tách ra do đó cải thiện trên các mô hình hiện có. một phép biến đổi được học, ở cấp độ pixel, từ một tên miền này sang tên miền khác, bằng cách sử dụng GAN. một phương pháp tiếp cận đối lập không giám sát đối với phân đoạn ngữ nghĩa, sử dụng cả kỹ thuật thích ứng miền cụ thể và toàn cầu, được đề xuất.

# **2. Dẫn nhập**

Trong ánh xạ không được giám sát, thuật toán học tập được cung cấp với các tập dữ liệu không có nhãn từ hai miền, A và B. Tập dữ liệu thứ nhất bao gồm các mẫu i.i.d từ phân phối ​​và tập dữ liệu thứ hai bao gồm các mẫu i.i.d từ phân phối.



Chính thức, với mục tiêu của chúng tôi là tìm hiểu hàm GAB, để ánh xạ các mẫu trong miền A tới các mẫu tương tự trong miền B, xem ví dụ bên dưới. Trong công việc trước đây [11, 28, 27], cần khôi phục đồng thời hàm GBA thứ hai, tương tự ánh xạ các mẫu trong miền B tới các mẫu tương tự trong miền A.

**Biện giải** Để cho phép học không giám sát của một ánh xạ định hướng, chúng tôi giới thiệu ràng buộc rằng các cặp đầu vào , ở khoảng cách nhất định với nhau, được ánh xạ tới cặp đầu ra GAB (x); GAB (x0) với khoảng cách tương tự, tức là



Figure 1: Mỗi bộ ba cho thấy hình ảnh túi xách nguồn, giày đích được tạo ra bởi GAB [G] và kết quả xấp xỉ GAB bằng một phép biến đổi tuyến tính phi định tính T, thu được mỗi điểm ảnh đầu ra dưới dạng tổ hợp tuyến tính của các pixel đầu vào. Việc chuyển đổi tuyến tính nắm bắt được bản chất của GAB cho thấy rằng phần lớn ánh xạ được thực hiện bằng một phép biến đổi không gian cố định.

và có tương quan cao. Như chúng tôi trình bày dưới đây, có thể giả định rằng ràng buộc này xấp xỉ trong nhiều kịch bản được chứng minh bởi công việc trước đây về dịch miền. Mặc dù gần đúng, nó là đủ, vì như được thể hiện trong [21], việc ánh xạ giữa các tên miền chỉ yêu cầu sự giám sát ít trên đầu của yêu cầu phân phối đầu ra của trình ánh xạ khớp với phân phối đích.

Xem xét, ví dụ, trường hợp ánh xạ giày tới các cạnh, như được trình bày trong Hình 4. Trong trường hợp này, các điểm cạnh chỉ đơn giản là một tập hợp con của tọa độ hình ảnh, được chọn bởi tiêu chí hình ảnh cục bộ. Nếu hình ảnh là trực quan tương tự như hình ảnh , có khả năng là các bản đồ cạnh của chúng tương tự nhau. Trong thực tế, sự tương tự này nhấn mạnh việc sử dụng thông tin gradient trong tài liệu tầm nhìn máy tính cổ điển. Vì vậy, trong khi khoảng cách được dự kiến ​​sẽ khác nhau trong hai tên miền, người ta có thể mong đợi một mối tương quan cao.

Tiếp theo, hãy xem xét trường hợp của túi xách để ánh xạ giày (Hình 4). Tương tự có xu hướng có cùng phân phối màu sắc hình ảnh trong các hình ảnh khác nhau. Giả sử rằng vị trí pixel không gian của túi xách theo phân phối chặt chẽ (tức là, tập hợp các hình ảnh túi xách có cùng hình dạng) và cùng lưu giữ cho giày, sau đó tồn tại một tập hợp các trường chuyển dịch hợp với chuyển đổi túi xách thành giày. Nếu có một chuyển vị, đó sẽ là một hoán vị cố định của các vị trí pixel, khoảng cách sẽ được bảo toàn. Trong thực tế, việc chuyển đổi hình ảnh phức tạp hơn.

Để nghiên cứu xem mô hình chuyển dịch hình ảnh có gần đúng hay không, chúng ta đã học được một phép biến đổi tuyến tính phi âm tính *T*  bản đồ đó, một kênh tại một thời điểm, hình ảnh túi xách có kích thước 64 × 64 × 3 cho hình ảnh giày đầu ra có cùng kích thước được đưa ra bởi phương pháp CycleGAN. Các cột của *T* có thể được hiểu là trọng số xác định sự lan truyền của khối lượng trong hình ảnh đầu ra cho mỗi vị trí pixel trong hình ảnh đầu vào. Nó được ước tính bằng cách giảm thiểu sai số bình phương của ánh xạ mọi kênh (R, G hoặc B) của một hình ảnh túi xách đến cùng một kênh trong giày phù hợp. Việc tối ưu hóa được thực hiện bằng độ dốc gốc với phép chiếu tới không gian của các ma trận không âm, nghĩa là, đánh số các phần tử âm của *T* tại mỗi lần lặp.

Các ánh xạ mẫu của ma trận *T* được thể hiện trong hình 1. Như có thể thấy, phép biến đổi tuyến tính không âm tính xấp xỉ CNN GAB nhiều lớp của CycleGAN ở một mức độ nào đó. Kiểm tra các phần tử của *T*, chúng chia sẻ một số thuộc tính với hoán vị: tổng trung bình của các hàng là 1:06 (SD 0:08) và 99: 5% các phần tử dưới 0:01.

Trong trường hợp thêm kính hoặc thay đổi giới tính hoặc màu tóc (Hình 3), một sửa đổi hình ảnh tương đối nhỏ, không thay đổi đáng kể phần lớn thông tin hình ảnh, đủ để tạo hiệu ứng hình ảnh mong muốn. Thay đổi như vậy có khả năng duy trì phần lớn khoảng cách hình ảnh hai chiều trước và sau khi chuyển đổi.

Trong trường hợp đầu máy tính được tạo ra ở các góc khác nhau so với xe xoay, được trình bày trong [11], khoảng cách tương quan cao một phần vì khu vực được chụp bởi đối tượng tiền cảnh là một chỉ báo tốt về tiếng kêu của vật thể. Khi ánh xạ giữa ngựa với ngựa vằn [28], kết cấu của da ngựa được chuyển thành hình ngựa vằn. Trong trường hợp này, hầu hết các thông tin hình ảnh không bị ảnh hưởng và phần được thay đổi được sửa đổi bởi một kết cấu đồng nhất, một lần nữa khoảng cách duy trì khoảng cách hai chiều. Trong hình 2 (a), chúng tôi so sánh khoảng cách *L1* trong không gian RGB của các cặp hình ảnh con ngựa với khoảng cách của các mẫu sau khi ánh xạ bởi mạng CycleGAN [28] được thực hiện, sử dụng việc thực hiện công khai. Rõ ràng là sự tương quan giữa các miền giữa các khoảng cách cặp đôi là rất cao. Chúng tôi cũng xem xét hình ảnh Cityscapes và các cặp nhãn chân lý trong Hình 2 (c), và thấy rằng có sự tương quan cao giữa khoảng cách. Đây cũng là trường hợp trong nhiều ánh xạ dựa trên văn học khác giữa các tập dữ liệu mà chúng ta đã thử nghiệm và các cặp chân lý cơ sở.

Mặc dù có ít nhược điểm khi làm việc với các cặp hình ảnh đào tạo so với làm việc với các hình ảnh đơn lẻ, để nghiên cứu thêm lượng thông tin cần thiết để căn chỉnh thành công, chúng tôi cũng xem xét khoảng cách giữa hai nửa của cùng một hình ảnh.

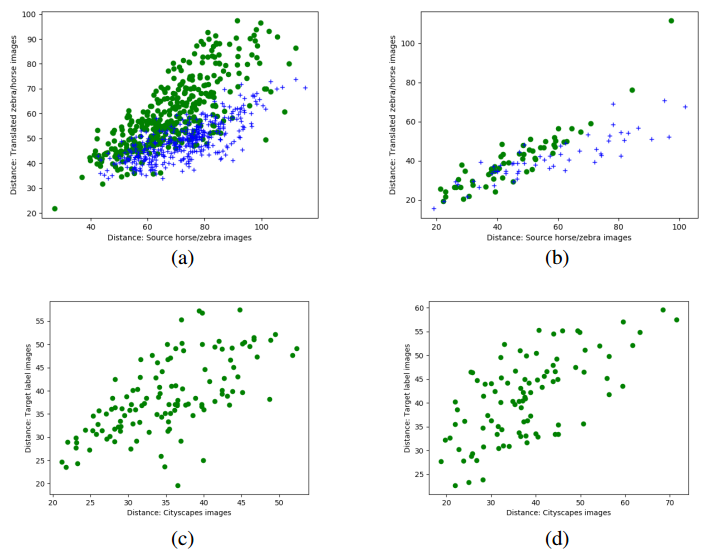


Figure 2: Xác định mối tương quan cao giữa khoảng cách trong các miền khác nhau. (a) Sử dụng mô hình CycleGAN [28], chúng tôi lập bản đồ ngựa cho ngựa vằn và ngược lại. Vòng tròn màu xanh lá cây được sử dụng cho khoảng cách giữa hai hình ảnh ngựa ngẫu nhiên và hai hình ảnh ngựa vằn được dịch tương ứng. Dấu thập màu xanh dương là hướng ngược lại dịch ngựa vằn thành hình ngựa. Tương quan Pearson cho dịch ngựa vằn là 0,777 (giá trị p 1: 7e − 113) và cho ngựa vằn với ngựa là 0,73 (giá trị p 8: 0e − 96). (b) Như trong (a) nhưng sử dụng khoảng cách giữa hai nửa của cùng một hình ảnh hoặc là một hình ảnh ngựa được dịch sang một con ngựa vằn hoặc ngược lại. Tương quan Pearson cho dịch ngựa vằn là 0,91 (giá trị p 9: 5e − 23) và cho ngựa vằn với ngựa là 0:87 (giá trị p 9: 7e − 19). (c) Hình ảnh thành phố và nhãn được liên kết. Vòng tròn màu xanh lá cây được sử dụng cho khoảng cách giữa hai hình ảnh thành phố và hai hình ảnh thật tương ứng. Tương quan Pearson là 0:65 (giá trị p 6: 0e − 16). (d) Như trong (c) nhưng sử dụng khoảng cách giữa hai nửa của cùng một hình ảnh. Tương quan Pearson là 0:65 (giá trị p 1: 4e − 12)

Chúng tôi so sánh khoảng cách L1 giữa nửa trái và nửa phải như được tính trên hình ảnh đầu vào đến hình ảnh được tạo trên hình ảnh được tạo ra hoặc hình ảnh thật tương ứng. Hình 2 (b) và hình 2 (d) trình bày kết quả cho ngựa để dịch ngựa vằn và cho cặp hình ảnh và nhãn hiệu Cityscapes tương ứng. Như có thể thấy, sự tương quan cũng rất quan trọng trong trường hợp này.

**Từ các mối tương quan với tổng số khác biệt tuyệt đối** Chúng tôi đã cung cấp sự biện minh và thực nghiệm bằng chứng cho nhiều ánh xạ ngữ nghĩa, có một mức độ tương quan cao giữa khoảng cách hai chiều trong hai miền. Nói cách khác, hãy cho là một vectơ của khoảng cách được chuẩn hóa theo đơn vị trung tâm và đơn vị chuẩn hóa trong một miền và cho là vectơ của khoảng cách chuẩn hóa thu được trong miền khác bằng cách dịch từng hình ảnh ra khỏi mỗi cặp giữa các tên miền, sau đó nên cao. Khi đào tạo GAB bản đồ, giá trị trung bình và phương sai được sử dụng để chuẩn hóa mỗi tên miền được ước tính trước dựa trên các mẫu đào tạo trong mỗi tên miền, mà giả định rằng phân bố bản đồ của các mẫu tương tự như phân phối đào tạo.

Khoảng cách hai chiều trong *dk*miền nguồn được cố định và tối đa hóa gây ra cặp đôi khoảng cách *dk* với giá trị tuyệt đối lớn để thống trị tối ưu hóa. Thay vào đó, chúng tôi đề xuất giảm thiểu tổng số chênh lệch tuyệt đối , điều này sẽ làm sai lệch sai số trong khoảng cách thống nhất. Hai tổn - và có liên quan cao và mối tương quan nghịch giữa chúng được tính toán rõ ràng cho các bản phân phối đơn giản và được hiển thị rất mạnh [1].

# **3. Các ràng buộc không được giám sát trên Ánh xạ đã học**

Có một số loại ràng buộc được đề xuất trong tài liệu, không yêu cầu các mẫu ghép nối. Đầu tiên, người ta có thể thực thi việc phân phối, mà chúng ta biểu thị là , để không thể phân biệt được với . Ngoài ra, người ta có thể yêu cầu ánh xạ từ A đến B và trở lại sẽ dẫn đến một bản đồ nhận dạng. Một ràng buộc khác được đề xuất, là cho mỗi . Chúng tôi xem xét các ràng buộc này và sau đó trình bày các ràng buộc mới mà chúng tôi đề xuất.

**Ràng buộc đối nghịch** Bộ đào tạo của chúng tôi được xem là hai bản phân phối riêng biệt Avà B được lấy mẫu từ phân phối miền nguồn và đích và , tương ứng. Đối với mạng GAB đã học, sự giống nhau giữa các phân phối và được mô hình hóa bởi một GAN. Điều này liên quan đến việc đào tạo một mạng lưới phân biệt đối xử . Sự mất mát được đưa ra bởi:



Mất mát này được giảm thiểu trên và được tối đa hóa trên . Khi cả và đều đã học đồng thời, có một biểu thức tương tự B,A), trong đó vai trò chuyển đổi A và B và hai tổn thất (và bốn mạng) được tối ưu hóa cùng nhau.

**Ràng buộc về tuần hoàn** Trong ba báo cáo gần đây [11, 28, 27], mất mát tuần hoàn được giới thiệu để dịch hình ảnh. Lý do là đưa ra một mẫu từ miền A, dịch nó sang miền B và sau đó quay lại miền A sẽ dẫn đến mẫu giống hệt nhau. Mất mát sau đây được thêm vào:



Chỉ tiêu *L1* được sử dụng ở trên đã được tìm thấy là chủ yếu là thích hợp hơn, mặc dù *L2* cho kết quả tương tự. Vì mất mát tuần hoàn đòi hỏi sự phục hồi của ánh xạ theo cả hai hướng, nó thường được sử dụng đối xứng, bằng cách xem xét A) + B)

Ràng buộc tuần hoàn thường được xem như là một yêu cầu nhất định cho các hàm có thể chấp nhận GAB và GBA. Tuy nhiên, giống như các ràng buộc dựa trên khoảng cách, nó là một sự gần đúng. Để thấy điều này, hãy xem xét ví dụ về ngựa vằn với ngựa. Lập bản đồ một con ngựa vằn với một con ngựa có nghĩa là mất các sọc. Do đó, bản đồ nghịch đảo không thể được mong đợi để phục hồi các sọc đầu vào chính xác.

**Danh tính mục tiêu tên miền** Một ràng buộc đã được sử dụng trong [22] và trong một số thí nghiệm trong [28] nói rằng *GAB* áp dụng cho các mẫu từ miền *B* thực hiện ánh xạ danh tính. Chúng tôi đã không thử nghiệm với ràng buộc này và nó được đưa ra ở đây để hoàn thành:



**Ràng buộc khoảng cách** Việc mất đối thủ đảm bảo rằng các mẫu từ phân phối *A* được dịch sang các mẫu trong phân phối *B*. Tuy nhiên, có rất nhiều ánh xạ có thể xảy ra. Với một ánh xạ cho *n* mẫu từ *A* đến *n* mẫu *B*, ta có thể xem xét bất kỳ hoán vị nào của các mẫu trong B như một ánh xạ hợp lệ và do đó, không gian ánh xạ hàm từ A đến B là rất lớn. Thêm ràng buộc tròn, thực thi ánh xạ từ *B* đến *A* là nghịch đảo của hoán vị xảy ra từ A đến B, làm giảm số lượng hoán vị được chấp nhận.

Để tiếp tục giảm không gian này, chúng tôi đề xuất một bản đồ bảo tồn khoảng cách, tức là khoảng cách giữa hai mẫu trong A phải được bảo toàn trong ánh xạ tới B. Do đó chúng tôi xem xét sự mất mát sau đây, đó là kỳ vọng về sự khác biệt tuyệt đối giữa khoảng cách trong mỗi tên miền lên đến quy mô:



nơi là các phương tiện (độ lệch chuẩn) của khoảng cách cặp đôi trong các tập huấn luyện từ *A* và *B*, tương ứng, và được ước tính trước.

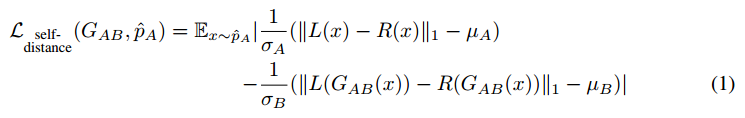
Trong thực tế, chúng tôi tính toán sự mất mát trên các cặp mẫu thuộc cùng một minibatch trong quá trình đào tạo. Ngay cả đối với minibatches với 64 mẫu, như trong DiscoGAN [11], xem xét tất cả các cặp là khả thi. Nếu cần thiết, đối với các lô nhỏ hơn, người ta có thể làm mẫu cho các cặp.

Khi hai ánh xạ đồng thời được học, B) được xác định tương tự. Trong cả hai trường hợp, sự khác biệt tuyệt đối về khoảng cách L1 giữa các cặp trong hai tên miền được xem xét.

So với tính tròn, giới hạn dựa trên khoảng cách không bị ảnh hưởng bởi sự cố thu gọn mô hình được mô tả trong [11]. Trong hiện tượng này, hai mẫu khác nhau từ miền A được ánh xạ tới cùng một mẫu trong miền B. Ánh xạ theo hướng ngược lại sau đó tạo ra trung bình của hai mẫu ban đầu, vì mẫu trong miền B nên được ánh xạ trở lại cả mẫu ban đầu và mẫu thứ hai trong A. Ràng buộc khoảng cách theo cặp ngăn cản điều này xảy ra.

**Ràng buộc tự khoảng cách** Có hay không ràng buộc khoảng cách có hiệu quả hơn so với ràng buộc tuần hoàn trong việc khôi phục sự liên kết, ràng buộc dựa trên khoảng cách có lợi thế là một mặt. Tuy nhiên, nó đòi hỏi rằng các cặp mẫu được chuyển giao cùng một lúc, trong đó, trong khi có ít ý nghĩa về quá trình đào tạo như nó hiện đang thực hiện, có thể ảnh hưởng đến khả năng thực hiện học trực tuyến. Hơn nữa, việc thực hiện chính thức CycleGAN [28] sử dụng minibatches kích thước một. Do đó, chúng tôi đề xuất một ràng buộc bổ sung, sử dụng một mẫu tại một thời điểm và so sánh khoảng cách giữa hai phần của cùng một mẫu.

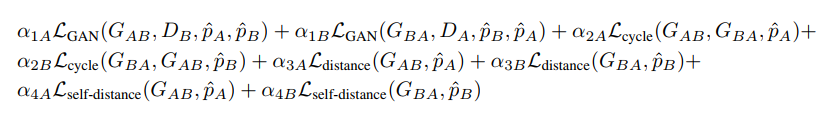
Cho là các toán tử đưa ra một hình ảnh đầu vào trả về phần bên trái hoặc bên phải của nó. Chúng tôi xác định sự mất mát sau:



trong đó và là độ lệch chuẩn và trung bình của khoảng cách hai chiều giữa hai một nửa của hình ảnh trong tập huấn luyện từ miền A, và tương tự cho và , ví dụ, với tập huấn luyện được ước tính trước là .

## **3.1 Network Architecture and Training (Kiến trúc và đào tạo mạng)**

Khi đào tạo mạng GAB; GBA; DB và DA, chúng tôi sử dụng khoản lỗ sau, được giảm thiểu tối đa trên GAB và GBA và tối đa hóa trên DB và DA:

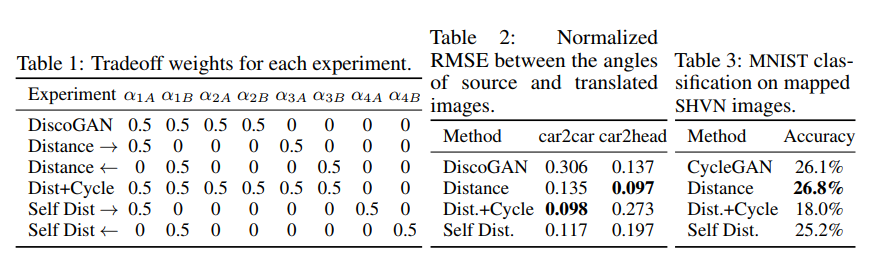


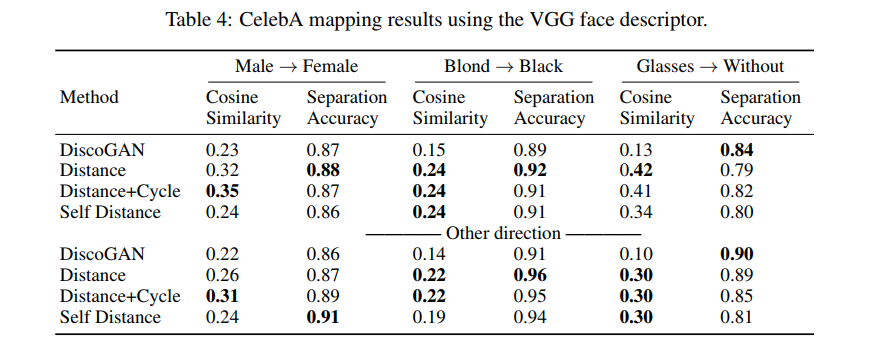
trong đó αiA, αiB là các tham số giao dịch. Chúng tôi không kiểm tra ràng buộc khoảng cách và ràng buộc khoảng cách tự, vì vậy trong mọi thử nghiệm, α3A = α3B = 0 hoặc α4A = α4A = 0. Khi thực hiện ánh xạ một mặt từ A đến B, chỉ α1A và α3A hoặc α4A không phải là số không.

Chúng tôi xem A và B là một tập hợp con của R3×s×s hình ảnh trong đó s là 64, 128 hoặc 256, tùy thuộc vào độ phân giải hình ảnh. Để so sánh trực tiếp kết quả của chúng tôi với công việc trước đó và sử dụng đường cơ sở mạnh nhất trong mỗi tập dữ liệu, chúng tôi sử dụng các kiến trúc máy phát và phán đoán của cả hai DiscoGAN và CycleGAN.

Trong DiscoGAN, trình tạo máy phát được xây dựng từ bộ giải mã bộ mã hóa. Bộ mã hóa bao gồm các lớp chập chững với bộ lọc 4 × 4 tiếp theo là các đơn vị kích hoạt Leaky ReLU. Bộ giải mã bao gồm các lớp giải mã với các bộ lọc 4 × 4 theo sau là một đơn vị kích hoạt ReLU. Sigmoid được sử dụng cho lớp đầu ra và chuẩn hóa hàng loạt được sử dụng trước khi kích hoạt ReLU hoặc Leaky ReLU. Từ 4 đến 5 lớp co giãn / giải mã được sử dụng, tùy thuộc vào các miền được sử dụng trong A và B (chúng tôi khớp với kiến trúc mã đã xuất bản cho mỗi tập dữ liệu). Bộ phán đoán tương tự như bộ mã hóa, nhưng có thêm một lớp chập chững như lớp đầu tiên và một đơn vị đầu ra sigmoid.

Kiến trúc CycleGAN cho máy phát điện dựa trên [10]. Các máy phát bao gồm hai lớp co giãn 2 bước, giữa 6 đến 9 khối còn lại tùy thuộc vào độ phân giải của hình ảnh và hai chuyển động được phân chia theo phần với bước sóng 1 = 2. Sự bình thường hóa thể hiện được sử dụng như trong [10]. Người phán đoán sử dụng 70 × 70 PatchGAN. Đối với đào tạo, CycleGAN sử dụng hai kỹ thuật bổ sung. Đầu tiên là thay thế khả năng nhật ký âm bằng cách mất ít nhất và thứ hai là sử dụng lịch sử hình ảnh cho các nhà phán đoán, thay vì đó chỉ là hình ảnh cuối cùng được tạo ra.





# **4. Experiments (Thử nghiệm)**

Chúng tôi so sánh nhiều phương pháp: đường cơ sở của DiscoGAN hoặc CycleGAN; ánh xạ một mặt sử dụng Ldistance (A -> B hoặc B -> A); sự kết hợp của phương pháp cơ bản với Ldistance; phương pháp tự cách khoảng. Đối với DiscoGAN, chúng tôi sử dụng cấu hình trọng lượng cố định cho tất cả thử nghiệm, như được hiển thị trong Tab. Đối với CycleGAN, có nhiều nhạy cảm hơn với các tham số và trong khi mẫu chung được bảo toàn, chúng tôi sử dụng trọng số khác nhau cho giới hạn khoảng cách tùy thuộc vào thử nghiệm, chữ số hoặc ngựa và ngựa vằn.

**Mô hình dựa trên DiscoGAN:**

Các tập dữ liệu đã được kiểm tra bởi DiscoGAN được đánh giá ở đây bằng cách sử dụng kiến trúc này. Trong các thử nghiệm ban đầu, CycleGAN không cạnh tranh trên những chiếc hộp này. Bộ thí nghiệm đầu tiên lập bản đồ hình ảnh ô tô xoay quanh ô tô hoặc người đứng đầu. Bộ dữ liệu xe 3D bao gồm hình ảnh được hiển thị của ô tô 3D có mức độ khác nhau trong khoảng thời gian 15o. Tương tự, tập dữ liệu đầu, bao gồm hình ảnh 3D của đầu xoay thay đổi từ −70o đến 70o. Đối với thử nghiệm car2car, bộ dữ liệu xe hơi được chia thành hai phần, một trong số đó được sử dụng cho A và một cho B (Nó được chia tách thành bộ huấn luyện và kiểm tra). Vì góc quay thể hiện nguồn biến thiên lớn nhất, và do phép quay được chia sẻ giữa các tập dữ liệu, chúng ta mong đợi nó là bất biến chính mà mạng học được, tức là ánh xạ ngữ nghĩa sẽ bảo toàn các góc.

Một thuật hồi quy đã được huấn luyện để tính toán góc của một hình ảnh xe nhất định dựa trên dữ liệu huấn luyện. Tab 2 hiển thị Lỗi trung bình gốc (RMSE) giữa góc của hình ảnh nguồn và hình ảnh đã dịch. Như có thể thấy, ánh xạ dựa trên khoảng cách cặp đôi dẫn đến lỗi thấp hơn so với bản đồ DiscoGAN, kết hợp cả hai cải thiện thêm kết quả và khoảng cách vượt trội hơn cả DiscoGAN và khoảng cách hai chiều. Việc thực hiện DiscoGAN ban đầu đã được sử dụng, nhưng do sự khác biệt về đánh giá (các biến hồi quy khác nhau) những con số này không tương thích với biểu đồ được hiển thị trong DiscoGAN.

Đối với car2head, giải pháp của DiscoGAN tạo ra hình ảnh phản chiếu và sự kết hợp ràng buộc tròn của DiscoGAN với giới hạn khoảng cách tạo ra giải pháp được xoay 90o. Chúng tôi coi những thành kiến này là sự mơ hồ trong việc lập bản đồ và không phải là sai lầm, và do đó, loại bỏ lỗi trung bình trước khi tính toán RMSE. Trong thử nghiệm này, khoảng cách hoạt động tốt hơn tất cả các phương pháp khác. Sự kết hợp của cả hai phương pháp là ít cạnh tranh hơn cả hai, có lẽ vì mỗi phương pháp kéo về phía một giải pháp khác nhau. Khoảng cách tự, tệ hơn thông tư trong tập dữ liệu này.

Một tập hợp các thí nghiệm khác phát sinh từ việc xem xét các hình ảnh khuôn mặt có và không có một thuộc tính nhất định. CelebA được chú thích cho nhiều thuộc tính bao gồm giới tính, màu tóc và sự tồn tại của kính trong hình ảnh. Sau đây chúng ta thực hiện ánh xạ giữa hai giá trị của mỗi thuộc tính trong ba thuộc tính này. Các kết quả được thể hiện trong tài liệu bổ sung với một số ví dụ trong Hình 3. Rõ ràng là phương pháp DiscoGAN (sử dụng các tác giả chưa sửa đổi) trình bày nhiều trường hợp lỗi hơn phương pháp dựa trên cặp của chúng ta. Phương pháp tự khoảng cách được thực hiện với nửa hình ảnh trên cùng và dưới cùng, thay vì từ trái sang phải, vì khuôn mặt là đối xứng. Phương pháp này cũng có vẻ tốt hơn DiscoGAN.

Để đánh giá khả năng dịch khuôn mặt tốt như thế nào, chúng tôi sử dụng lớp biểu diễn của VGG đối mặt trên hình ảnh trong A và đầu ra của nó trong B. Người ta có thể giả định rằng hai hình ảnh phù hợp sẽ có nhiều tính năng tương tự và do đó biểu diễn VGG sẽ tương tự. Các điểm tương đồng cosin, như được đánh giá giữa các hình ảnh đầu vào và các phiên bản được ánh xạ của chúng, được hiển thị trong Tab. 4. Trong mọi trường hợp, khoảng cách ghép đôi tạo ra nhiều khuôn mặt đầu vào-đầu ra tương tự hơn. Tự thực hiện khoảng cách hơi tồi tệ hơn so với các cặp, nhưng thường tốt hơn DiscoGAN. Áp dụng vòng tròn cùng với khoảng cách cặp, cung cấp kết quả tốt nhất nhưng đòi hỏi, không giống như khoảng cách, học cả hai bên cùng một lúc.

Mặc dù chúng tôi tạo ra hình ảnh phù hợp hơn với chỉ số mô tả khuôn mặt, khả năng tạo hình ảnh trung thành với phân phối thứ hai của chúng tôi không bị suy giảm. Điều này được chứng minh bằng cách học một phân loại tuyến tính giữa hai tên miền dựa trên các mẫu đào tạo và sau đó áp dụng nó cho một tập hợp các hình ảnh thử nghiệm trước và sau khi ánh xạ. Độ chính xác phân tách giữa hình ảnh thử nghiệm đầu vào và phiên bản được ánh xạ cũng được hiển thị trong Tab. 4. Như có thể thấy, khả năng tách của phương thức của chúng ta tương tự như của DiscoGAN (nó phát sinh từ các thuật ngữ GAN được chia sẻ).

Chúng tôi cũng thực hiện nghiên cứu người dùng để đánh giá chất lượng kết quả của chúng tôi. Người dùng lần đầu tiên được trình bày với một tập hợp các hình ảnh thực từ tập dữ liệu. Sau đó, 50 cặp hình ảnh ngẫu nhiên được trình bày cho người dùng trong một giây, một hình ảnh được đào tạo bằng cách sử dụng DiscoGAN và một hình ảnh sử dụng phương pháp của chúng tôi. Người dùng được yêu cầu quyết định hình ảnh nào trông thực tế hơn. Bài kiểm tra được thực hiện trên 22 người dùng. Trên giày để dịch túi xách, bản dịch của chúng tôi hoạt động tốt hơn trên 65% các trường hợp. Đối với túi xách cho giày, điểm số là 87%. Từ nam sang nữ, cả hai phương pháp đều cho thấy một điểm số thực tế tương tự (51% đến 49% của DiscoGAN). Do đó, chúng tôi đã hỏi một câu hỏi thứ hai: với khuôn mặt của một người đàn ông, cái mà trong số hai biến thể nữ được tạo ra là phù hợp hơn với khuôn mặt nguyên thủy. Phương pháp của chúng tôi thắng 88% thời gian.

Ngoài ra, trong tài liệu bổ sung, chúng tôi so sánh các tổn thất của phân biệt đối xử GAN đối với các phương pháp khác nhau và cho thấy rằng các giá trị này gần giống nhau. Chúng tôi cũng đo lường các tổn thất của các phương pháp khác nhau trong quá trình thử nghiệm, ngay cả khi chúng không được tối ưu hóa trực tiếp. Ví dụ, mặc dù các ràng buộc này không được thực thi, các phương pháp dựa trên khoảng cách dường như trình bày một sự mất mát theo chu kỳ thấp, trong khi DiscoGAN trình bày một tổn thất khoảng cách tương đối cao hơn.

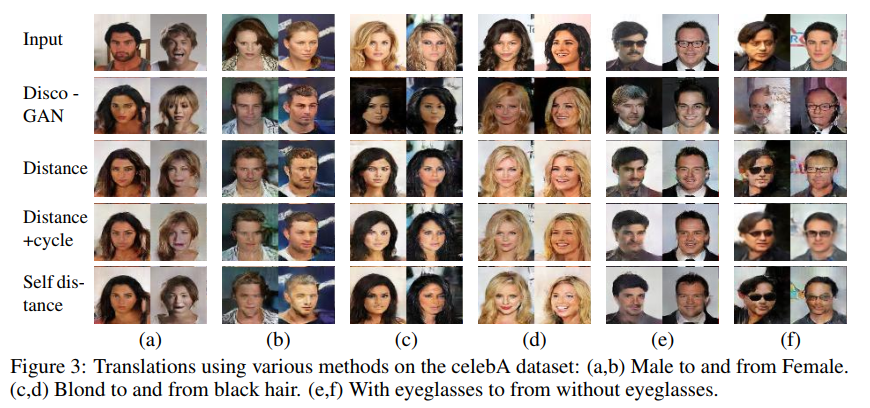
Kết quả mẫu của việc vẽ bản đồ giày đến túi xách và mép giày và ngược lại bằng cách sử dụng kiến trúc cơ bản của DiscoGAN được thể hiện trong Hình 3. Các kết quả khác được thể hiện trong phần bổ sung. Nhìn chung, kết quả của cách tiếp cận dựa trên khoảng cách dường như tốt hơn sau đó DiscoGAN trong khi kết quả của tự khoảng cách là hơi tồi tệ hơn. Sự kết hợp giữa DiscoGAN và khoảng cách thường hoạt động tốt nhất.

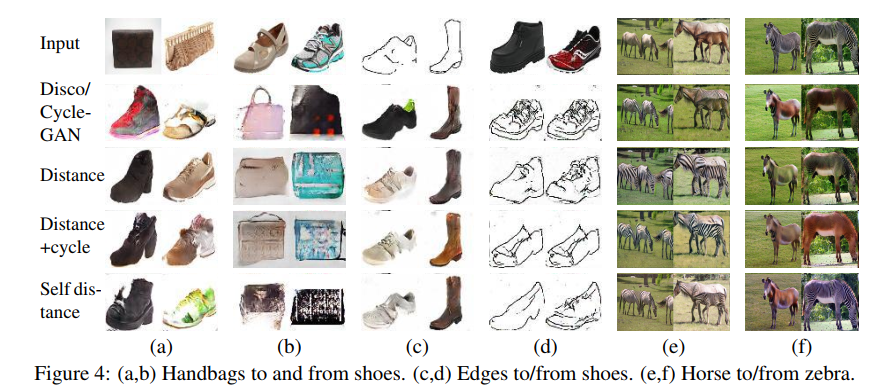
**Mô hình dựa trên CycleGAN:**

Sử dụng kiến trúc CycleGAN chúng tôi lập bản đồ ngựa và ngựa vằn, xem Hình 4 và tài liệu bổ sung cho các ví dụ. Lưu ý rằng trên ngựa vằn ánh xạ với ngựa, tất cả các phương pháp không thành công mặc dù theo những cách khác nhau. Chủ quan, có vẻ như phương pháp khoảng cách + chu kỳ cho thấy lời hứa nhất trong bản dịch này.

Để có được kết quả bằng số, chúng tôi sử dụng phương pháp CycleGAN cơ bản cũng như các phương pháp của chúng tôi để dịch từ Số nhà phố (SVHN) sang MNIST. Độ chính xác sau đó được đo bằng không gian MNIST bằng cách sử dụng mạng neural được huấn luyện cho nhiệm vụ này. Kết quả được hiển thị trong Tab. 3 và trực quan trong Bổ sung. Trong khi phương pháp dựa trên khoảng cách hai chiều được cải thiện theo phương pháp cơ sở, vẫn còn một khoảng cách lớn giữa cài đặt không giám sát và bán giám sát được trình bày trong [22], đạt được kết quả cao hơn nhiều. Điều này có thể được giải thích bởi số lượng lớn thông tin không liên quan trong các hình ảnh SVHN (các ví dụ được thể hiện trong phần bổ sung). Kết hợp khoảng cách dựa trên khoảng cách với vòng tròn không hoạt động tốt trên tập dữ liệu này.

Chúng tôi cũng thực hiện đánh giá định tính bằng cách sử dụng điểm FCN như trong [28]. Chỉ số FCN đánh giá hình ảnh khả năng tương tác bằng cách chụp ảnh cảnh quan thành phố được tạo và tạo nhãn bằng thuật toán phân đoạn ngữ nghĩa. Nhãn được tạo sau đó có thể được so sánh với nhãn sự thật trên mặt đất. Kết quả FCN được đưa ra dưới dạng ba số đo: độ chính xác trên mỗi pixel, độ chính xác cho mỗi lớp và IOU lớp. Phương pháp GAN khoảng cách của chúng tôi là thích hợp hơn trên cả ba điểm (0,53 so với 0,52, 0,19 so với 0,17 và 0,11 so với 0,11, tương ứng). Giá trị p kiểm tra t được ghép nối lần lượt là 0:29, 0: 002 và 0:42. Trong một nghiên cứu người dùng tương tự như nghiên cứu của DiscoGAN ở trên, bản dịch thành phố của chúng tôi đạt 71% cho sự thật khi so sánh với CycleGAN. Khi nhìn vào sự tương đồng với hình ảnh chân lý, chúng tôi đạt 68%.





# **5. Conclusion (Kết luận)**

Chúng tôi đã đề xuất một mất mát dựa trên khoảng cách không giám sát để tìm hiểu một ánh xạ đơn (không có nghịch đảo của nó), mà theo kinh nghiệm thực hiện tốt hơn sự mất mát theo chu kỳ. Thật thú vị khi lưu ý rằng sự mất mát mới được áp dụng cho các giá trị hình ảnh RGB thô. Điều này trái ngược với tất cả công việc mà chúng tôi biết là tính toán sự giống nhau về hình ảnh. Rõ ràng, mô tả hình ảnh hoặc kích hoạt mạng lớp thấp có thể được sử dụng. Tuy nhiên, bằng cách xem xét chỉ các giá trị RGB, chúng tôi không chỉ hiển thị tiện ích chung của phương pháp của chúng tôi, mà còn chứng minh thêm rằng một lượng thông tin tối thiểu là cần thiết để tạo thành sự tương tự giữa hai miền liên quan.