

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH



BÁO CÁO ĐỒ ÁN

**Khảo sát Image Stiching và ứng dụng
ghép hình ảnh đa góc nhìn**

Môn học : Xử lý ảnh & video số

Giảng viên : PGS.TS Lý Quốc Ngọc

Ths Phạm Thanh Tùng

Ths Nguyễn Mạnh Hùng

Nhóm : The Fourth

25 - 11 - 2024, Thành Phố Hồ Chí Minh

Mục lục

1 Thông tin thành viên	3
2 Đóng góp của bài khảo sát	3
3 Giới thiệu chủ đề	3
3.1 Động lực nghiên cứu	3
3.1.1 Ý nghĩa khoa học	3
3.1.2 Ý nghĩa ứng dụng	4
3.2 Đặt vấn đề	4
3.3 Phát biểu bài toán	4
3.3.1 Framework chung	5
3.4 Thách thức	5
4 Các công trình nghiên cứu liên quan	6
4.1 Phương pháp truyền thống	6
4.1.1 Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images	7
4.1.2 Combination of feature-based and area-based image registration technique for high resolution remote sensing image	9
4.1.3 Image registration with Fourier-based image correlation: a comprehensive review of developments and applications	13
4.1.4 Bảng so sánh các phương pháp truyền thống	16
4.2 Phương pháp học sâu	17
4.2.1 Efficient Unsupervised Image Stitching Using Attention Mechanism with Deep Homography Estimation	17
4.2.2 Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images	19
4.2.3 Implicit Neural Image Stitching With Enhanced and Blended Feature Reconstruction	26
4.2.4 Bảng so sánh các phương pháp học sâu	32
5 Phương pháp	33
5.1 Bộ dữ liệu huấn luyện	33
5.2 Phương pháp truyền thống	35
5.3 Phương pháp học sâu	37
5.3.1 Kiến trúc mô hình	38

5.3.2	Quá trình huấn luyện	42
6	Cài đặt và thử nghiệm	44
6.1	Cấu hình môi trường	44
6.2	Phương pháp truyền thống	44
6.3	Phương pháp học sâu	45
7	Kết quả và đánh giá	47
7.1	Các thang đo	47
7.2	Phương pháp truyền thống	48
7.2.1	Kết quả thử nghiệm	48
7.2.2	Cài đặt thang đo	52
7.2.3	Kết quả thang đo	53
7.3	Phương pháp học sâu	55
7.3.1	Kết quả thử nghiệm	55
7.3.2	Cài đặt thang đo	59
7.3.3	Kết quả thang đo	60
7.4	So sánh giữa hai phương pháp	61
7.4.1	Về giá trị PSNR	61
7.4.2	Về giá trị SSIM	61
7.4.3	Phân tích dựa trên ảnh kết quả	61
7.4.4	Ưu nhược điểm của từng phương pháp	62
7.4.5	Kết luận	62
8	Kết luận và hướng phát triển	62
8.1	Kết luận	62
8.2	Hạn chế	63
8.3	Hướng phát triển tương lai	63
9	Tài liệu tham khảo	65

1 Thông tin thành viên

MSSV	Họ tên	Lớp	Email
22127384	Dương Quang Thắng	22TGMT	dqthang22@clc.fitus.edu.vn
22127385	Nguyễn Quốc Thắng	22TGMT	nqthang22@clc.fitus.edu.vn
22127141	Ngô Hoàng Nam Hưng	22TGMT	nhnhung22@clc.fitus.edu.vn
22127301	Nguyễn Gia Nguyễn	22TGMT	ngnguyen22@clc.fitus.edu.vn

2 Đóng góp của bài khảo sát

Ghép ảnh là một lĩnh vực nghiên cứu quan trọng trong thị giác máy tính và đồ họa máy tính, nhằm giải quyết những hạn chế về trường quan sát bằng cách kết hợp hai hoặc nhiều ảnh chồng lên nhau để tạo thành một bức ảnh toàn cảnh có góc nhìn rộng và độ phân giải cao. Công nghệ này được ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm hình ảnh phong cảnh và kiến trúc, hình ảnh vệ tinh và bản đồ, hình ảnh y tế, thực tế ảo, và an ninh giám sát. Bài khảo sát này cung cấp cái nhìn toàn diện về các thành phần chính của quy trình ghép ảnh, trình bày chi tiết các phương pháp tiếp cận, bao gồm các giải pháp truyền thống và các giải pháp dựa trên học sâu. Đối với hướng tiếp cận truyền thống, bài khảo sát trình bày những phương pháp đã từng được sử dụng bao gồm dựa trên vùng và dựa trên đặc trưng. Bên cạnh đó, chúng tôi tập trung nghiên cứu các tiến bộ đạt được trên nền tảng các kỹ thuật trích xuất đặc trưng truyền thống như SIFT, SURF, ORB, đồng thời phân tích và ứng dụng các phương pháp hiện đại dựa trên mô hình học sâu để cải thiện khả năng trích xuất và biểu diễn đặc trưng. Bài khảo sát thảo luận về ưu nhược điểm của từng phương pháp, cùng với các thử nghiệm đánh giá và so sánh trên các tập dữ liệu tiêu chuẩn. Cuối cùng, bài khảo sát đề cập đến những thách thức hiện tại trong ghép ảnh và thảo luận các hướng phát triển tiềm năng trong tương lai.

3 Giới thiệu chủ đề

3.1 Động lực nghiên cứu

3.1.1 Ý nghĩa khoa học

Nghiên cứu về ghép ảnh có ý nghĩa khoa học nằm ở khả năng giúp mở rộng góc nhìn và thu thập thông tin, với mục tiêu tổng hợp từ nhiều hình ảnh riêng rẽ có sự chồng lấn. Kỹ thuật này thúc đẩy sự phát triển của các thuật toán mạnh để xử lý, căn chỉnh và kết hợp hình ảnh nhằm tối đa hóa tính chính xác và độ chi tiết trên bức ảnh. Nghiên cứu về ghép ảnh góp phần vào sự hiểu biết sâu hơn về các thuật toán tối ưu hóa và các phương pháp trích xuất đặc trưng, các phép biến đổi hình học cũng như các phương pháp xử lý ảnh. Việc cải thiện độ chính xác và tốc độ xử lý mang lại những lợi ích không chỉ cho lĩnh vực thị giác máy tính mà còn cho các lĩnh vực khác như học máy, học sâu, đồ họa và các hệ thống thông minh.

3.1.2 Ý nghĩa ứng dụng

Về mặt ứng dụng, ghép ảnh trở thành công cụ kỹ thuật quan trọng trong nhiều lĩnh vực:

- **Y tế:** việc tạo ra hình ảnh tổng hợp từ những ảnh cắt MRI, CT hay X-ray giúp các chuyên gia y tế có cái nhìn toàn diện hơn về cấu trúc và tình trạng các cơ quan trong cơ thể. Điều này hỗ trợ trong việc chuẩn đoán và lập phác đồ điều trị hiệu quả hơn.
- **Địa lý và bản đồ:** được ứng dụng trong ghép ảnh vệ tinh để tạo ra các bản đồ có độ phân giải cao phục vụ nghiên cứu địa lý, phân tích sự biến đổi của môi trường và phân tích thảm họa thiên nhiên.
- **Robotics và thực tế ảo:** cho phép hệ thống thị giác của robot tự hành và máy bay không người lái có tầm nhìn rộng bao quát giúp điều hướng hành vi. Ghép ảnh cũng được dùng trong phát triển các ứng dụng thực tế ảo (VR) và thực tế ảo tăng cường (AR) giúp tăng trải nghiệm người dùng bằng việc tạo ra môi trường chân thực và đa góc nhìn.
- **Giải trí và nhiếp ảnh:** trong nhiếp ảnh và làm phim, ghép ảnh tạo ra các ảnh toàn cảnh và video với trường quan sát rộng. Trong các phần mềm chỉnh sửa ảnh tích hợp ghép ảnh để tạo ra hình ảnh chất lượng cao từ nhiều tấm ảnh phục vụ thương mại như Adobe Photoshop hay AutoStitch.

3.2 Đặt vấn đề

Trong nhiều ứng dụng thực tế như y tế, địa lý, nhiếp ảnh, và thị giác máy tính, có nhiều tác vụ yêu cầu việc ghi nhận và xử lý dữ liệu hình ảnh với độ chi tiết cao. Tuy nhiên, do giới hạn về góc nhìn và độ phân giải của các thiết bị ghi hình, việc chụp toàn cảnh hoặc chụp chi tiết một khu vực lớn thường phải thực hiện qua nhiều hình ảnh riêng lẻ. Điều này đặt ra một thách thức lớn trong việc kết hợp những hình ảnh này lại với nhau để tạo ra một ảnh toàn cảnh liền mạch và có chất lượng cao.

Mục tiêu của nghiên cứu và phát triển kỹ thuật ghép ảnh là xây dựng các phương pháp và thuật toán để tự động căn chỉnh và ghép ảnh với độ chính xác và phân giải cao. Những thuật toán này cần đảm bảo rằng ảnh tổng hợp không chỉ liền mạch về mặt hình học mà còn hài hòa về mặt màu sắc và chi tiết.

3.3 Phát biểu bài toán

Bài toán yêu cầu ghép nhiều hình ảnh có sự chồng lấp một phần, được chụp từ các góc nhìn hoặc vị trí khác nhau, thành một hình ảnh duy nhất biểu diễn cảnh toàn thể dưới một góc nhìn thống nhất.

Bao gồm việc xác định và áp dụng các phép biến đổi hình học để căn chỉnh các hình ảnh sao cho chúng khớp với nhau về không gian và xử lý các vùng chồng lấp để tạo ra kết quả liền mạch, tự nhiên và nhất quán về cả cấu trúc hình học lẫn đặc tính thị giác.

Đầu vào là tập hợp gồm nhiều hình ảnh số. Trong đó:

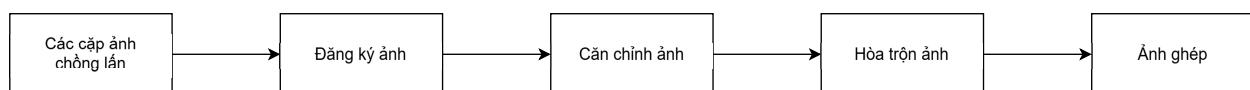
- Các hình ảnh này phải có sự chồng lấn nhất định và được chụp từ nhiều góc nhìn khác nhau của cùng một cảnh hoặc đối tượng.

- Các hình ảnh có thể có các sai lệch về hình học do yếu tố phối cảnh hoặc sự dịch chuyển thiết bị.

Đầu ra là một hình ảnh tổng hợp với yêu cầu:

- Hình ảnh tổng hợp có độ phân giải cao, các thành phần cần được căn chỉnh chính xác với sự kết nối về màu sắc và cường độ sáng nhằm tránh sự đứt đoạn giữa các vùng nối rõ rệt.
- Hình ảnh cần đảm bảo sự nhất quán về mặt hình học giữa các ảnh.

3.3.1 Framework chung



Như được thể hiện trong Hình 1, quy trình ghép ảnh bao gồm các bước xử lý hình ảnh khác nhau: đăng ký, căn chỉnh và hòa trộn.

- **Đăng ký hình ảnh** là việc thiết lập sự tương ứng dữ liệu giữa nhiều cặp ảnh mô tả cùng một cảnh. Mục tiêu của đăng ký hình ảnh là căn chỉnh các hình ảnh sao cho các điểm tương ứng giữa chúng khớp chính xác.
- **Căn chỉnh hình ảnh** đề cập đến việc sau khi đăng ký hình ảnh, các phép biến đổi hình học được tính toán để làm biến dạng và hợp nhất các vùng chồng lấn của các ảnh. Đây là bước các ảnh được điều chỉnh để khớp chính xác với nhau về mặt hình học, bao gồm xử lý các biến dạng và sai lệch phối cảnh.
- **Hòa trộn hình ảnh** nhằm làm mờ các đường nối giữa các ảnh, điều chỉnh độ sáng và màu sắc để tránh các hiện tượng không đồng nhất, tạo sự liền mạch và tự nhiên cho bức ảnh.

3.4 Thách thức

Một trong những vấn đề phổ biến nhất là hiện tượng bóng mờ hoặc hình ảnh kép) trong vùng chồng lấn giữa các ảnh. Hiện tượng này thường xảy ra khi có sự thay đổi trong cảnh chụp do các đối tượng di chuyển giữa các lần chụp. Ví dụ, khi ghép ảnh của một con đường đông đúc, các xe cộ hoặc người đi bộ xuất hiện tại các vị trí khác nhau trong các khung hình, dẫn đến sự không nhất quán khi các ảnh được xếp chồng lên nhau. Hiện tượng này không chỉ ảnh hưởng đến chất lượng hình ảnh tổng thể mà còn làm giảm tính chân thực của hình ảnh toàn cảnh.

Một thách thức khác là biến dạng hình học trong vùng chồng lấn. Biến dạng này phát sinh khi các hình ảnh được chụp từ các góc độ khác nhau hoặc với các khoảng cách khác nhau đến các điểm trong cảnh. Kết quả là, các điểm tương đồng giữa các hình ảnh có thể bị dịch chuyển hoặc sai lệch về vị trí, gây khó khăn trong việc căn chỉnh và ghép nối chính xác. Hơn nữa, một phần cảnh trong một hình ảnh có thể bị che khuất bởi các đối tượng khác trong hình ảnh kia, làm mất đi các thông tin cần thiết để ghép nối chính xác.

Bên cạnh đó, sự khác biệt về độ sâu và thị sai cũng là một vấn đề phổ biến, đặc biệt trong các cảnh ba chiều phức tạp. Thị sai xảy ra khi các đối tượng trong cảnh có khoảng cách khác nhau so với máy ảnh, dẫn đến sự khác biệt về vị trí của chúng trong các ảnh đầu vào. Vấn đề này trở nên khó khăn hơn khi khoảng cách giữa máy ảnh và các đối tượng thay đổi đáng kể, hoặc khi có các vật thể gần và xa xuất hiện đồng thời trong cảnh.

Ngoài các yếu tố trên, các thách thức bổ sung như sự khác biệt về độ sáng) và màu sắc giữa các hình ảnh cũng cần được xem xét. Những thay đổi này thường do các điều kiện ánh sáng khác nhau hoặc do cài đặt máy ảnh không đồng nhất, dẫn đến sự không hài hòa trong kết quả ghép nối cuối cùng.

4 Các công trình nghiên cứu liên quan

4.1 Phương pháp truyền thống

Việc tìm kiếm sự tương đồng giữa những hình ảnh có sự chồng lấn là một bước quan trọng trong quá trình ghép ảnh, nhằm xác định cách thức để căn chỉnh và một cách chính xác. Điều này có thể được chia thành hai hướng tiếp cận bao gồm dựa trên vùng và dựa trên đặc trưng.

Đối với dựa trên vùng, phương pháp này tập trung vào việc so sánh và căn chỉnh các vùng ảnh (hoặc các pixel) giữa các ảnh cần ghép. Mục tiêu là xác định các vùng có sự tương đồng nhất trong ảnh. Đầu tiên, phương pháp này sử dụng phép biến đổi hình học (như biến đổi affine hoặc homography) để căn chỉnh các bức ảnh với nhau, sau đó tính toán sự sai lệch bằng cách đo lường sự khác biệt giữa các pixel của các ảnh khi đã áp dụng phép biến đổi. Sự sai lệch này được tính toán qua một hàm lỗi như sai lệch bình phương (SSD) hoặc sai lệch tuyệt đối (SAD), chúng ta cần thiểu hóa hàm lỗi này với mục tiêu làm cho các ảnh càng khớp càng tốt. Để tối thiểu hóa hàm lỗi, các phương pháp tối ưu hóa phép biến đổi được sử dụng bao gồm: Levenberg-Marquardt hoặc Gauss-Newton. Quá trình tối ưu này cần được lặp lại nhiều lần để đạt được kết quả tốt nhất. Phương pháp này hoạt động hiệu quả trong điều kiện lý tưởng như tập các ảnh có sự chồng lấn tốt (gần như hoàn toàn trùng khớp), không có sự thay đổi về góc nhìn (góc chụp, độ nghiêng, biến dạng hình học) và không phụ thuộc vào cường độ sáng, độ tương phản của các ảnh. Tuy nhiên, hiệu suất của phương pháp này trong thực tế rất thấp vì độ phức tạp tính toán và tài nguyên lớn, các phương pháp tối ưu hóa như Levenberg-Marquardt hoặc Gauss-Newton yêu cầu tính toán trên toàn bộ ảnh nhiều lần làm tăng đáng kể thời gian tính toán, đặc biệt là tập hình ảnh có độ phân giải cao. Cũng như gặp khó khăn khi xử lý tập ảnh có sự thay đổi về cường độ sáng, độ tương phản hoặc các biến dạng phức tạp (thay đổi về góc chụp, biến dạng hình học, độ nghiêng). Chính vì vậy phương pháp này cho thấy sự không ổn định trong môi trường thực tế.

Phương pháp dựa trên đặc trưng là phát hiện, mô tả và kết hợp các điểm đặc trưng tương ứng của ảnh, từ đó xác định mối quan hệ hình học giữa các ảnh để tạo ra một ảnh liền mạch. Đặc trưng là các điểm trong ảnh có tính nổi bật và duy nhất, ổn định trước các biến đổi hình học như xoay, thay đổi tỷ lệ, góc nhìn và cường độ ánh sáng. Các thuật toán như SIFT (Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Features) và ORB (Oriented FAST and Rotated BRIEF) được sử dụng để phát hiện các đặc trưng này. Mỗi đặc trưng sau đó được biểu diễn thông qua một vector đặc trưng mô tả các thuộc tính vùng xung quanh điểm, như hướng, độ sáng và kết cấu, nhằm làm cơ sở cho quá trình ghép nối, khi ghép nối, các vector đặc trưng được so sánh để tìm các cặp điểm tương ứng

giữa các ảnh. Để loại bỏ nhiễu, thuật toán RANSAC (Random Sample Consensus) được áp dụng, giúp loại bỏ các điểm ghép nối sai và chỉ giữ lại các điểm phù hợp cho phép biến đổi. Cuối cùng, phép biến đổi hình học như Homography được sử dụng để căn chỉnh ảnh dựa trên các điểm đã ghép nối, cho phép ghép nối chính xác các ảnh có góc nhìn khác nhau.

Cả hai phương pháp dựa trên vùng và dựa trên đặc trưng đều thuộc vào nhóm các phương pháp dựa trên miền không gian bởi đều thực hiện các phép tính toán và biến đổi trực tiếp trên pixel ảnh. Ngoài ra phương pháp dựa trên miền tần số cũng được đề cập, trong đó tập ảnh được chuyển từ miền không gian về miền tần số và sử dụng các phép biến đổi tần số để tìm kiếm sự tương đồng và căn chỉnh hình ảnh. Phương pháp này sử dụng biến đổi Fourier để chuyển đổi các giá trị pixel của ảnh thành tổ hợp các tần số. Cụ thể trong miền tần số, các tín hiệu ảnh được chuyển đổi và mô tả thông qua các thành phần tần số thay vì các giá trị pixel, mỗi thành phần tần số đại diện cho một phần thông tin chi tiết của ảnh ở các tần số khác nhau bao gồm tần số thấp (thường đại diện cho đặc trưng hoặc cấu trúc lớn như nền, hình dạng lớn) và tần số cao (thường đại diện cho các cấu trúc chi tiết như biên cạnh). Sau đó tính toán sự tương đồng của các thành phần tần số và thực hiện căn chỉnh sao cho độ tương đồng của các miền tần số là tối ưu. Phương pháp dựa trên miền tần số hoạt động hiệu quả khi tập ảnh có sự thay đổi về cường độ sáng và độ tương phản, độ biến dạng hình học thấp và tiết kiệm tài nguyên tính toán trong một số trường hợp, đặc biệt phương pháp này có thể phát hiện cấu trúc tổng thể của ảnh. Tuy nhiên phương pháp này tồn tại hạn chế đáng kể khi áp dụng cho tập ảnh có sự biến dạng hình học lớn, độ tương phản thấp, mờ và phức tạp trong việc xử lý các chi tiết nhỏ của ảnh. Ngoài ra hình ảnh kết quả có thể bị biến dạng khi chuyển từ miền tần số về miền không gian do sự mất thông tin khi xử lý các thành phần tần số, đôi khi các chi tiết quan trọng có thể bị làm mất, điều này dẫn đến việc ảnh ghép không chính xác hoặc không tự nhiên. Độ phức tạp tính toán cũng tăng lên khi xử lý các tập hình ảnh lớn gây tốn kém tài nguyên.

4.1.1 Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images

- **Tác giả:** Ebrahim Karami, Siva Prasad, Mohamed Shehata
- **Công bố tại:** Newfoundland Electrical and Computer Engineering St. John's, Canada, 2015

Nghiên cứu này tập trung so sánh hiệu năng của các thuật toán SIFT, SURF, BRIEF, và ORB trong việc ghép ảnh dưới các điều kiện biến dạng như thay đổi tỷ lệ, xoay, nhiễu, cắt xén, và biến dạng mắt cá. Các tiêu chí đánh giá gồm số lượng điểm đặc trưng phát hiện, tỷ lệ trùng khớp chính xác, và thời gian thực thi.

- **SIFT (Scale Invariant Feature Transform):** Được Lowe giới thiệu năm 2004, SIFT là thuật toán mạnh trong xử lý xoay, biến đổi affine và khác biệt góc nhìn. Quy trình bao gồm phát hiện điểm đặc trưng với DoG (Difference of Gaussian), tinh chỉnh vị trí điểm, gán hướng gradient, và tạo vector mô tả đặc trưng. Tuy nhiên, nhược điểm chính là thời gian tính toán dài, khó đáp ứng yêu cầu thời gian thực.
- **SURF (Speeded Up Robust Features):** Là phiên bản cải tiến của SIFT, SURF tối ưu hóa thời gian bằng cách thay thế DoG bằng bộ lọc hộp (box filter), kết hợp với ảnh tích phân và ma trận Hessian. Tính năng phân loại điểm dựa trên dấu hiệu Laplacian giúp tăng tốc quá trình so khớp.

- BRIEF (Binary Robust Independent Elementary Features):** Sử dụng mô tả đặc trưng dạng nhị phân, BRIEF có độ phức tạp thấp, phù hợp với các ứng dụng đòi hỏi tốc độ. Tuy nhiên, thuật toán này kém hiệu quả trong xử lý ảnh xoay hoặc biến đổi phức tạp.
- ORB (Oriented FAST and Rotated BRIEF):** Là sự kết hợp giữa FAST và BRIEF, ORB cải tiến khả năng bắt biến với xoay và tỷ lệ thông qua điều chỉnh hướng mô tả đặc trưng và sử dụng điểm Harris để chọn điểm đặc trưng tốt hơn. ORB đạt hiệu năng cao trong cả tốc độ và độ chính xác, phù hợp với các ứng dụng thời gian thực.

	Time (sec)	Kpnts1	Kpnts2	Matches	Match rate (%)
SIFT	0.16	248	260	166	65.4
SURF	0.03	162	271	110	50.8
ORB	0.03	261	423	158	46.2

Hình 2: Kết quả so sánh với cường độ màu sắc lớn

	Time (sec)	Kpnts1	Kpnts2	Matches	Match rate (%)
SIFT	0.16	248	260	166	65.4
SURF	0.03	162	271	110	50.8
ORB	0.03	261	423	158	46.2

Hình 3: Kết quả so sánh với ảnh bị xoay

Angle →	0	45	90	135	180	225	270
SIFT	100	65	93	67	92	65	93
SURF	99	51	99	52	96	51	95
ORB	100	46	97	46	100	46	97

Hình 4: Kết quả so sánh với các góc quay

	Time (sec)	Kpnts1	Kpnts2	Matches	Match rate (%)
SIFT	0.25	248	1210	232	31.8
SURF	0.08	162	581	136	36.6
ORB	0.02	261	471	181	49.5

Hình 5: Kết quả so sánh với biến dạng tỷ lệ

	Time (sec)	Kpnts 1	Kpnts 2	Matches	Match rate (%)
SIFT	0.133	248	229	150	62.89
SURF	0.049	162	214	111	59.04
ORB	0.026	261	298	145	51.88

Hình 6: Kết quả so sánh với biến dạng cắt

	Time (sec)	Kpnts 1	Kpnts 2	Matches	Match rate (%)
SIFT	0.132	248	236	143	59.09
SURF	0.036	162	224	85	44.04
ORB	0.012	261	282	125	46.04

Hình 7: Kết quả so sánh với biến dạng mắt cá

	Time (sec)	Kpnts1	Kpnts2	Matches	Match rate (%)
SIFT	0.115	248	242	132	53.8
SURF	0.059	162	385	108	39.48
ORB	0.027	261	308	155	54.48

Hình 8: Kết quả so sánh khi thêm biến dạng muối tiêu 30%

4.1.2 Combination of feature-based and area-based image registration technique for high resolution remote sensing image

- **Tác giả:** Gang Hong, Yun Zhang
- **Công bố tại:** Geoscience and Remote Sensing Symposium, 2007

a. **Động lực nghiên cứu:** Các phương pháp truyền thống thường không cân bằng được giữa tính chính xác, khả năng khử biến dạng và hiệu suất tính toán.

- *Phương pháp dựa trên vùng:* Tốt trong so sánh mức xám và ghép vùng lớn. Nhưng dễ bị ảnh hưởng bởi nhiều và khác biệt mức xám giữa ảnh cảm biến và ảnh tham chiếu. Kém hiệu quả khi đối mặt với các biến dạng cục bộ.
- *Phương pháp dựa trên đặc trưng:* Ổn định hơn trong môi trường có nhiều và biến dạng cục bộ, nhưng khó đạt độ chính xác cao hơn mức pixel. Cần thuật toán phức tạp để trích xuất đặc trưng.

⇒ Sự kết hợp tận dụng thế mạnh của cả hai, giúp giảm nhược điểm của từng phương pháp riêng lẻ.

b. Tập dữ liệu kiểm thử:

IKONOS và QuickBird là các vệ tinh viễn thám có độ phân giải cao. Dưới đây là thông tin chi tiết về hai tập dữ liệu:

• IKONOS:

- **Nguồn gốc:** IKONOS là vệ tinh thương mại viễn thám quang học đầu tiên có khả năng chụp ảnh độ phân giải cao, được phóng lên quỹ đạo vào ngày 24 tháng 9 năm 1999 bởi Space Imaging (nay là Maxar Technologies).

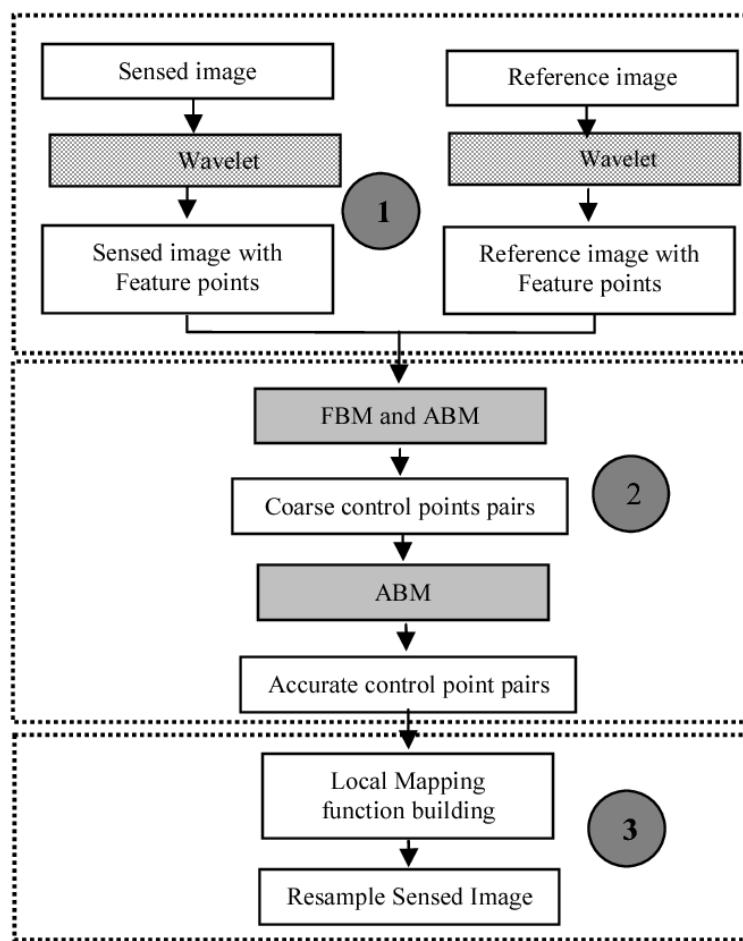
– **Đặc điểm nổi bật:**

- * *Độ phân giải không gian:*
 - Ảnh toàn sắc (panchromatic): 1 m.

- **Ảnh đa phổ (multispectral):** 4 m.
- * **Phạm vi quang phổ:**
 - Ảnh toàn sắc: 450–900 nm.
 - Ảnh đa phổ: Gồm các dải xanh, đỏ, lục và cận hồng ngoại.
- * **Phạm vi quét:** Lên tới 11 km x 11 km trên mặt đất.
- * **Ứng dụng:** Lập bản đồ đô thị, nông nghiệp, giám sát môi trường, quản lý tài nguyên, và an ninh.
- **Trạng thái:** Vệ tinh đã ngừng hoạt động từ năm 2015, nhưng dữ liệu IKONOS vẫn được sử dụng rộng rãi trong nghiên cứu và ứng dụng.
- **QuickBird:**
 - **Nguồn gốc:** QuickBird là vệ tinh thương mại độ phân giải cao của DigitalGlobe (hiện thuộc Maxar Technologies), được phóng lên quỹ đạo vào ngày 18 tháng 10 năm 2001.
 - **Đặc điểm nổi bật:**
 - * **Dộ phân giải không gian:**
 - Ảnh toàn sắc: Tối đa 0.61 m (khi chụp tại nadir).
 - Ảnh đa phổ: 2.44 m.
 - * **Phạm vi quang phổ:**
 - Ảnh toàn sắc: 450–900 nm.
 - Ảnh đa phổ: Gồm các dải xanh, đỏ, lục và cận hồng ngoại.
 - * **Phạm vi quét:** Lên tới 16.5 km trên mặt đất.
 - * **Ứng dụng:** Hỗ trợ lập bản đồ chính xác cao, phân tích sử dụng đất, quản lý tài nguyên thiên nhiên, quy hoạch đô thị và theo dõi thiên tai.
 - **Trạng thái:** Vệ tinh QuickBird đã ngừng hoạt động vào năm 2015, nhưng dữ liệu thu thập được vẫn có giá trị lớn trong nghiên cứu khoa học và ứng dụng thực tế.

c. Nguyên lý hoạt động:

Điểm nổi bật của nghiên cứu này so với các phương pháp truyền thống khác chính là cải tiến ở bước đăng ký hình ảnh. Phương pháp đề xuất tận dụng và kết hợp linh hoạt ưu điểm của hai kỹ thuật phổ biến: “Dựa trên đặc trưng” và “Dựa trên vùng”. Sự kết hợp này không chỉ tăng cường độ chính xác mà còn đảm bảo hiệu quả trong việc đăng ký ảnh dưới nhiều điều kiện biến dạng khác nhau.



Hình 9: Luồng hoạt động của quy trình Đăng ký hình ảnh

Phần 1. Xác định các điểm điều khiển tương ứng trong ảnh tham chiếu và ảnh cảm biến

Sử dụng phân tích đa phân giải với wavelet: Cả ảnh tham chiếu và ảnh cảm biến được phân tích thông qua kỹ thuật phân rã wavelet, tạo ra các ảnh pyramid từ mức thấp đến mức cao.

- *LL (Low Level):* Là ảnh xấp xỉ, chứa thông tin tổng quát của ảnh. Thành phần này được sử dụng để đăng ký ảnh vì nó bao quát thông tin chung nhất.
- *LH, HL, HH (Detailed images):* Các ảnh chi tiết tương ứng với các đặc trưng (như cạnh, góc) theo chiều ngang (LH), chiều dọc (HL), và chéo (HH). Từ các thành phần này, có thể tạo ra ảnh đặc trưng (*magnitude image*), hỗ trợ việc xác định các điểm điều khiển.

Sau khi trích xuất các điểm đặc trưng từ các thành phần chi tiết, phương pháp *relaxation* được áp dụng để tối thiểu hóa sai số giữa các điểm điều khiển trên hai ảnh.

Quá trình tối ưu hóa: Các điểm điều khiển thô được điều chỉnh qua tối ưu hóa để loại bỏ các điểm sai lệch do biến dạng hoặc nhiễu. Điều này giúp cải thiện độ chính xác của các cặp điểm điều khiển.

Phần 2. Tinh chỉnh các điểm điều khiển và tạo ra cặp điểm điều khiển chính xác.

Ghép ảnh dựa trên đặc trưng: Áp dụng các phương pháp như *cross-correlation matching* hoặc *probability relaxation method* để đối chiếu các đặc trưng giữa ảnh tham chiếu và ảnh cảm biến.

Tinh chỉnh các điểm điều khiển: Để tăng độ chính xác của các cặp điểm điều khiển:

- *Phương pháp Least Squares Matching (LSM):* Giúp giảm thiểu sai số bằng cách tìm ra các giá trị tối thiểu giữa các điểm tương ứng trong ảnh cảm biến và ảnh tham chiếu.
- *Phân bố điểm điều khiển đồng đều:* Ảnh được chia thành các lưới nhỏ, mỗi lưới chứa một số điểm điều khiển. Phương pháp này ngăn chặn sự tập trung quá nhiều điểm ở một khu vực, đảm bảo tính đại diện toàn bộ ảnh.

Phần 3. Xây dựng hàm ánh xạ và tái mẫu ảnh.

Xây dựng hàm ánh xạ: Hàm ánh xạ mô phỏng phép chuyển đổi giữa ảnh cảm biến và ảnh tham chiếu, có thể là:

- *Phép biến đổi affine:* Dùng để hiệu chỉnh sự thay đổi tuyến tính giữa hai ảnh.
- *Phép biến đổi projective:* Xử lý các biến dạng phức tạp hơn như phối cảnh.

Trong trường hợp ảnh cảm biến bị biến dạng địa hình, thay vì sử dụng một bộ hệ số duy nhất cho toàn bộ ảnh, ảnh sẽ được chia thành các vùng nhỏ. Mỗi vùng sử dụng hàm ánh xạ riêng, tăng độ chính xác khi ghép ảnh.

Tái mẫu ảnh (Resampling the Sensed Image): Sau khi xác định các hệ số ánh xạ, ảnh cảm biến được tái mẫu để khớp chính xác với ảnh tham chiếu. Quá trình này đảm bảo sự liên kết hoàn hảo giữa hai ảnh trong không gian chung.

d. Ưu điểm & Nhược điểm:

Ưu điểm:

- **Tự động xác định số lượng lớn các cặp điểm điều khiển:** Sử dụng kỹ thuật phân tích đa phân giải với wavelet cho phép trích xuất các điểm đặc trưng từ các thành phần chi tiết của ảnh (*LH*, *HL*, *HH*) một cách tự động. Điều này loại bỏ sự phụ thuộc vào việc chọn điểm thủ công, giúp tăng độ chính xác và hiệu quả của quá trình.
- **Tối ưu hóa thời gian tính toán:** Phân tích ảnh qua các cấp độ phân giải thấp đến cao (*pyramid image*) giúp giảm đáng kể khối lượng dữ liệu cần xử lý tại mỗi bước. Kỹ thuật này chia ảnh gốc thành các thành phần tương ứng với các dải tần số khác nhau, từ đó tối ưu hóa tài nguyên tính toán.
- **Hạn chế biến dạng cục bộ:** Phân vùng ảnh cảm biến thành các khu vực nhỏ và áp dụng hàm ánh xạ riêng cho từng vùng. Phương pháp này giúp giảm thiểu tác động của biến dạng cục bộ và cải thiện độ chính xác của quá trình ghép ảnh.

Nhược điểm:

- **Yêu cầu thời gian thiết lập ngưỡng cho việc chọn đặc trưng:** Mặc dù phương pháp tự động hóa trong việc chọn điểm điều khiển, việc xác định ngưỡng tối ưu để trích xuất các điểm đặc trưng từ ảnh tham chiếu và ảnh cảm biến đòi hỏi thời gian. Nếu ngưỡng không được thiết lập hợp lý, có thể dẫn đến việc chọn sai điểm, ảnh hưởng đến chất lượng kết quả.
- **Thời gian xử lý khi tối ưu hóa các điểm điều khiển:** Sử dụng phương pháp *relaxation* để giảm sai số giữa các điểm điều khiển mang lại độ chính xác cao, nhưng quá trình này có thể tiêu tốn nhiều thời gian, đặc biệt khi số lượng điểm điều khiển lớn hoặc khi ảnh gấp biến dạng phức tạp.

4.1.3 Image registration with Fourier-based image correlation: a comprehensive review of developments and applications

- **Tác giả:** Xiaohua Tong, Zhen Ye, Yusheng Xu, Sa Gao, Huan Xie, Qian Du, Shijie Liu, Xiong Xu, Sicong Liu, Kuifeng Luan, Uwe Stilla
- **Công bố tại:** IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019

Bài báo này cung cấp một cái nhìn tổng quan toàn diện về kỹ thuật tương quan ảnh dựa trên biến đổi Fourier (Fourier-based image correlation), một phương pháp mạnh mẽ và được ứng dụng rộng rãi trong việc căn chỉnh ảnh (image registration). Bài báo tập trung vào ba khía cạnh chính: nguyên tắc cơ bản, các phương pháp tính toán mức độ dịch chuyển subpixel, và các ứng dụng thực tế.

1. Giới thiệu về Fourier-Based Image Correlation: Đây là kỹ thuật căn chỉnh ảnh dựa trên việc phân tích sự tương quan trong miền tần số. Nó khai thác tính chất dịch chuyển của biến đổi Fourier: sự dịch chuyển trong miền không gian tương ứng với sự thay đổi pha tuyến tính trong miền tần số.

Ưu điểm:

- **Hiệu quả:** Tính toán nhanh chóng nhờ sử dụng biến đổi Fourier nhanh (FFT).
- **Chính xác:** Có khả năng đạt độ chính xác subpixel.
- **Mạnh mẽ:** Ít bị ảnh hưởng bởi nhiễu, đặc biệt là nhiễu tần số cao và sự thay đổi cường độ sáng, nhờ tập trung vào thông tin pha.

Hai phương pháp chính:

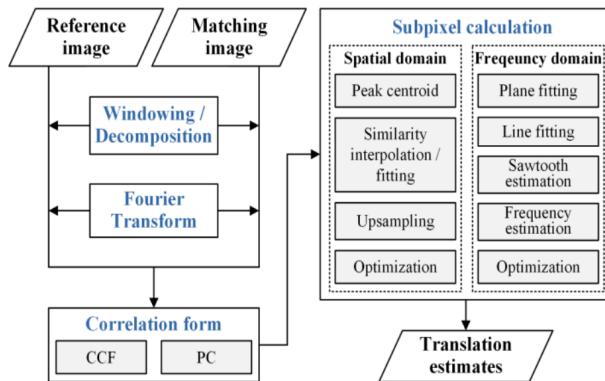
- **Cross-Correlation in the Frequency Domain (CCF):** Tính toán tương quan chéo giữa hai ảnh trong miền tần số.
- **Phase Correlation (PC):** Chỉ sử dụng thông tin pha, loại bỏ ảnh hưởng của biến độ, giúp tăng độ mạnh mẽ với nhiễu và sự thay đổi cường độ sáng. PC tạo ra một đỉnh tương quan sắc nét, giúp xác định chính xác độ dịch chuyển.

Mở rộng với Fourier-Mellin Transform (FM): Để xử lý các trường hợp ảnh bị xoay và thay đổi tỷ lệ, biến đổi FM được sử dụng. FM biến đổi ảnh sang không gian log-polar,

nơi mà phép xoay và thay đổi tỷ lệ trở thành phép dịch chuyển, từ đó có thể ước lượng được các tham số này.

2. Các phương pháp tính toán Subpixel:

Để đạt được độ chính xác cao hơn mức pixel, bài báo đi sâu vào các phương pháp tính toán subpixel, phân thành hai nhóm:



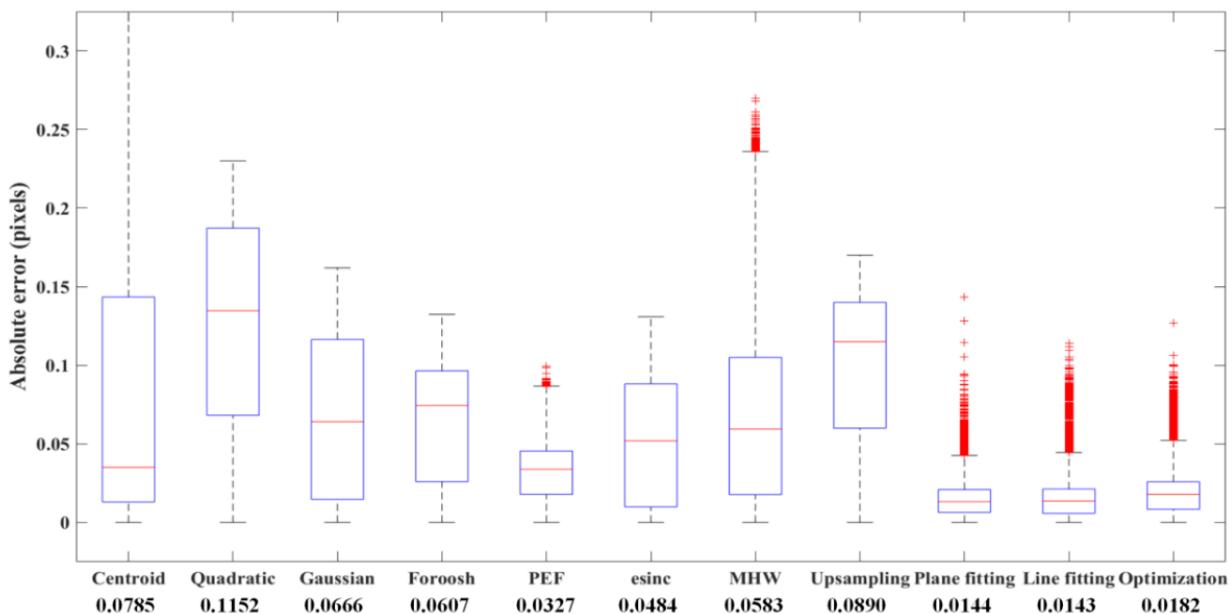
Hình 10: Quy trình làm việc của phương pháp tương quan hình ảnh Fourier-based dựa trên subpixel

Nhóm 1: Tính toán trong miền không gian (Spatial Domain):

- **Nguyên lý:** Dựa trên việc nội suy hoặc khớp bề mặt tương quan (correlation surface) xung quanh vị trí cực đại (peak) đã tìm được ở bước pixel-level.
- **Các phương pháp tiêu biểu:**
 - **Peak Centroid:** Tính trọng tâm của vùng lân cận xung quanh peak. Đơn giản nhưng dễ bị systematic error. (Hình 4 cho thấy độ chính xác thấp)
 - **Quadratic/Gaussian/sinc/Dirichlet/Modified sinc/MHW fitting:** Khớp bề mặt tương quan với các hàm toán học (parabol, Gaussian, sinc, v.v.). Cải thiện độ chính xác so với centroid. Trong đó, Gaussian và Dirichlet fitting cho kết quả tốt hơn. (Xem Bảng I và Hình 4).
 - **Upsampling:** Tăng độ phân giải của ảnh hoặc cross-power spectrum trước khi tính toán. Tốn kém về mặt tính toán và bộ nhớ, đặc biệt với ảnh lớn.
- **Ưu điểm:** Thường nhanh hơn các phương pháp trong miền tần số.
- **Nhược điểm:** Kém chính xác hơn, nhạy cảm với nhiễu hơn.

Nhóm 2: Tính toán trong miền tần số (Frequency Domain):

- **Nguyên lý:** Khai thác trực tiếp thông tin pha trong miền tần số để ước tính độ dịch chuyển subpixel.
- **Các phương pháp tiêu biểu:**
 - **Plane/Line fitting:** Khớp pha của cross-power spectrum với một mặt phẳng hoặc khớp các hàng/cột của ma trận hiệu pha (phase difference matrix) với các đường thẳng. Có thể kết hợp với kỹ thuật SVD để giảm nhiễu.



Hình 11: Biểu đồ MAE so sánh độ chính xác các phương pháp

- **Sawtooth signal estimation:** Xem ma trận hiệu pha như tín hiệu răng cưa 2D và ước tính tần số của nó.
- **Frequency estimation:** Ước tính tần số tức thời của cross-power spectrum.
- **Optimization:** Cực tiểu hóa hàm mục tiêu dựa trên sự khác biệt giữa cross-power spectrum đo được và lý thuyết (ví dụ: Frobenius norm).

- **Ưu điểm:** Thường chính xác hơn, ít bị ảnh hưởng bởi nhiễu hơn (đặc biệt là nhiễu tần số cao).
- **Nhược điểm:** Thường chậm hơn các phương pháp trong miền không gian, do không cần bước biến đổi ngược.

3. Ứng dụng

Bài báo tổng kết các ứng dụng chính của Fourier-based image correlation trong ba lĩnh vực:

- **Mở rộng các thuật toán image registration:** Cung cấp ước lượng ban đầu cho các thuật toán lặp, hoặc kết hợp với các phương pháp khác để nâng cao độ chính xác và hiệu quả.
- **Xử lý ảnh và tín hiệu tần số:** Nhận dạng sinh trắc học, xử lý địa vật lý, xử lý ảnh và video, ước lượng chuyển động.
- **Xử lý ảnh viễn thám:** Căn chỉnh ảnh, đo đặc biến dạng bề mặt, phân tích sự thay đổi theo thời gian.

4. Thách thức và hướng phát triển:

- **Cân bằng giữa độ chính xác và hiệu quả tính toán:** Cần phát triển các phương pháp subpixel vừa chính xác vừa nhanh, đặc biệt là cho dữ liệu ảnh lớn.

- Xử lý ảnh có kích thước cửa sổ nhỏ:** Cần cải thiện hiệu suất của các phương pháp Fourier-based với kích thước cửa sổ nhỏ, nơi mà nguyên lý bất định Heisenberg có thể ảnh hưởng đến kết quả.
- Phát triển các framework phù hợp với từng ứng dụng:** Cần tối ưu hóa các framework, bao gồm các bước tiền xử lý, hậu xử lý, và chiến lược matching, cho các ứng dụng cụ thể.
- So sánh và đánh giá định lượng:** Cần có các bộ dữ liệu chuẩn và các phương pháp đánh giá để so sánh hiệu suất của các phương pháp khác nhau một cách khách quan.

4.1.4 Bảng so sánh các phương pháp truyền thống

Tiêu chí	Feature-Based: Image Matching Using SIFT, SURF, BRIEF and ORB	Combination of Feature-Based and Area-Based Registration	Image Registration with Fourier-Based Image Correlation
Đầu vào	Hình ảnh gốc với các biến đổi như xoay, tỉ lệ, nhiễu, fisheye, shear.	Ảnh vệ tinh có độ phân giải cao (từ các vệ tinh như IKONOS và QuickBird).	Các ảnh có sự khác nhau như độ phân giải, góc quay, nhiễu hoặc biến dạng
Đầu ra	Số lượng điểm đặc trưng, tỷ lệ khớp, thời gian thực thi cho từng thuật toán.	Ảnh đã được đăng ký (aligned) với độ chính xác cao, giảm thiểu sai lệch do biến dạng hình học và điều kiện ánh sáng khác biệt.	Ảnh hoàn chỉnh với các chỉ số đánh giá độ chính xác, độ mượt, hiệu suất, tỷ lệ khớp điểm
Tập dữ liệu	Các hình ảnh với các biến đổi (xoay, tỉ lệ, nhiễu, fisheye, shear) để đánh giá các thuật toán.	Dữ liệu ảnh từ hai vệ tinh IKONOS và QuickBird.	Dữ liệu ảnh phổ biến với các biến đổi, độ lệch, tịnh tiến, xoay, tỉ lệ

Hình 12: Bảng so sánh 1

Đăng ký hình ảnh	Phương pháp dựa trên đặc trưng sử dụng các điểm đặc trưng (keypoints) để tìm mối tương quan giữa các hình ảnh, hỗ trợ xác định phép biến đổi cần thiết.	Kết hợp hai phương pháp: 1. Dựa trên đặc trưng (FBM): Trích xuất đặc trưng thông qua phân tích wavelet. 2. Dựa trên vùng (ABM): Ghép nối đặc trưng bằng cross-correlation và relaxation.	Sử dụng biến đổi Fourier để tính độ tương quan giữa các ảnh Không trích xuất đặc trưng mà sử dụng thông tin toàn cục trong không gian tần số, giúp giảm tác động nhiễu hay thay đổi ánh sáng
Căn chỉnh hình ảnh	Dựa trên các phép biến đổi affine hoặc phi tuyến, các hình ảnh được căn chỉnh để các vùng tương ứng khớp chính xác trên cùng một hệ tọa độ.	Sử dụng các điểm điều khiển tự động xác định qua trích xuất wavelet và tinh chỉnh bằng phương pháp Least Squares Matching (LSM).	Ước tính độ dịch chuyển subpixel sử dụng tương quan pha
Hòa trộn hình ảnh	Sau khi căn chỉnh, các hình ảnh được kết hợp mượt mà bằng cách tối ưu vùng giao thoa để giảm các đường biên hoặc hiện tượng bóng mờ.	Tái mẫu (resampling) dựa trên hàm ánh xạ (mapping function) được xây dựng từ các điểm điều khiển chính xác.	Sử dụng kỹ thuật như Feathering để giảm chênh lệch vùng biên. Áp dụng Multiband Blending để đảm bảo mượt mà và chi tiết cao.
Đánh giá	SIFT, SURF và ORB, hiệu quả trong xử lý biến đổi hình ảnh như xoay, tỉ lệ, và affine, với độ chính xác cao nhưng hạn chế trên ảnh ít chi tiết hoặc nhiễu;	Giảm biến dạng cục bộ và cải thiện độ chính xác tự động mà không cần lựa chọn thủ công. Tuy nhiên, cần thời gian để tối ưu hóa điểm điều khiển và thiết lập ngưỡng lý tưởng cho các đặc trưng.	Có tiềm năng độ chính xác cao với phương pháp subpixel tính độ dịch chuyển ít bị ảnh hưởng bởi nhiễu và độ tương phản 36

Hình 13: Bảng so sánh 2

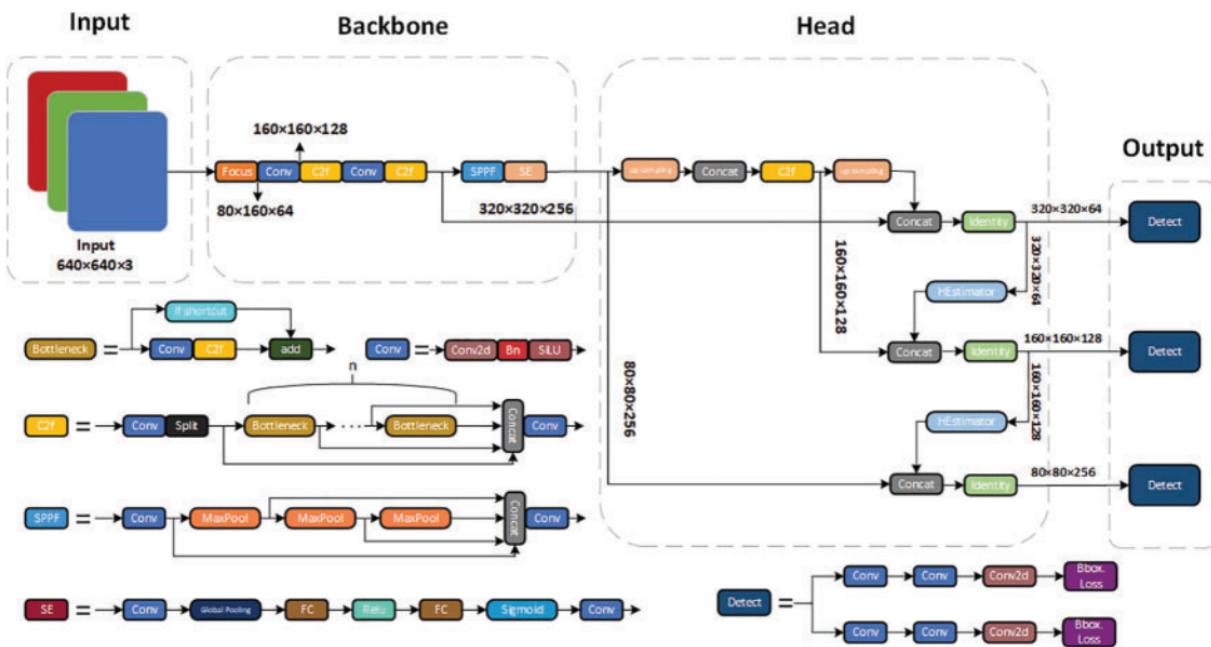
4.2 Phương pháp học sâu

Với học sâu, thay vì dựa vào các kỹ thuật ghép nối truyền thống, các mô hình học sâu học cách phát hiện và kết hợp hình ảnh từ dữ liệu đầu vào. Các phương pháp này sử dụng các mạng thần kinh hoặc các kiến trúc mô hình để xử lý và ghép nối hình ảnh.

4.2.1 Efficient Unsupervised Image Stitching Using Attention Mechanism with Deep Homography Estimation

- Tác giả:** Chubin Qin, Xiaotian Ran
- Công bố tại:** Tech Science Press, 2024

Nghiên cứu này nhận thấy rằng các kỹ thuật ghép ảnh dựa trên đặc trưng thường gặp khó khăn khi xử lý những hình ảnh có ít đặc trưng đáng kể hoặc bị suy giảm chất lượng nghiêm trọng. Đồng thời, sự thiếu hụt các bộ dữ liệu gán nhãn đầy đủ về các cảnh trong thực tế làm hạn chế hiệu quả của các phương pháp học có giám sát. Để khắc phục các vấn đề này, tác giả đã đề xuất một phương pháp ghép ảnh không giám sát, dựa trên framework YOL0v8 (You Only Look Once version 8), tích hợp mạng học sâu và cơ chế chú ý. Phương pháp này có các cải tiến chính như sau:



Hình 14: Kiến trúc YOLO được sử dụng

- **Cải tiến Backbone:** Backbone được tái cấu trúc với các thành phần cải tiến:
 - *Tách Detection Head:* Chia thành hai phần riêng biệt: phân loại (Binary Cross-Entropy) và hồi quy (CIOU + Distribution Focal Loss), giúp tăng cường khả năng phát hiện đặc trưng nhỏ và cải thiện độ chú ý vào vùng ảnh độ phân giải thấp.
 - *Thay đổi module:* Module C3 được thay thế bằng C2f, tích hợp Efficient Layer Aggregation Networks (ELAN) và cải tiến Feature Fusion với Path Aggregation Network (PAN) và Feature Pyramid Network (FPN).
 - *Xử lý biến dạng phi tuyến:* Sử dụng lõi tam giác để giảm thiểu biến dạng phi tuyến và tích hợp Spatial Pyramid Pooling-Fast (SPPF) cho xử lý đa tỷ lệ.
- **Cải tiến Head:** Cấu trúc Head cũ được loại bỏ và thay bằng mạng đồng nhất để ước lượng biến đổi đồng nhất (*deep homography estimation network*). Các cải tiến gồm:
 - Sử dụng padding để lấp đầy các pixel không hợp lệ.
 - Áp dụng phương pháp *ablation* nhằm loại bỏ nội dung không cần thiết ở các vùng pixel không hợp lệ.
 - Tối ưu hóa quá trình ghép ảnh toàn diện, giảm thiểu lỗi khớp đặc trưng (*mismatch*) và hiện tượng bóng ma (*ghosting artifacts*).
- **Cải tiến hàm mất mát (loss function):** Một mặt nạ (*mask*) được tạo bằng cách ngưỡng hóa giá trị pixel, với các giá trị lớn hơn 1 được coi là hợp lệ. Hàm mất mát được tính toán dựa trên sai số trung bình tuyệt đối (MAE) ở từng cấp độ, sau đó tổng hợp các thành phần mất mát với trọng số tương ứng.

Nghiên cứu này tận dụng bộ dữ liệu Warped COCO cho giai đoạn tiền huấn luyện, sau khi đã được tinh chỉnh để đảm bảo độ ổn định. Quá trình huấn luyện và kiểm thử mô hình được thực hiện trên hai bộ dữ liệu chính:

- **UDIS:** Bao gồm 10,500 mẫu dành cho huấn luyện và 1,100 mẫu dành cho kiểm thử. Tất cả các mẫu đều được chuẩn hóa về độ phân giải 640×640 pixels.
- **MVS-Synth:** Bao gồm 6,200 mẫu dùng để huấn luyện và 150 mẫu dùng để kiểm thử, với độ phân giải đồng nhất là 640×640 pixels.

Các bộ dữ liệu được lựa chọn nhằm đảm bảo sự đa dạng và tính nhất quán trong việc đánh giá hiệu quả của mô hình.

Dataset	Indicators	Homo	APAP [15]	UDIS-RSFI [8]	Our model
UDIS [8]	PSNR (\uparrow)	21.25	21.84	23.80	26.34
	SSIM (\uparrow)	0.7105	0.6952	0.7929	0.8414
	PSNR (\uparrow)	17.80	21.25	24.56	26.42
	SSIM (\uparrow)	0.6308	0.8434	0.8345	0.8494
MVS-Synth [35]	Model size (MB)	–	–	2105	188.3
	GFLOPs	–	–	10.5	14.5

Hình 15: So sánh thử nghiệm của PSNR và SSIM trên từng mô hình

Dataset	Indicators	YOLOv5	YOLOv8	YOLOv8 + A	YOLOv8 + B	Our model (A + B)
UDIS [8]	PSNR (\uparrow)	22.23	22.42	23.13	25.42	26.34
	SSIM (\uparrow)	0.7305	0.7329	0.7529	0.8103	0.8414
	PSNR (\uparrow)	22.12	22.13	23.36	25.11	26.42
	SSIM (\uparrow)	0.7408	0.7483	0.7645	0.8173	0.8494
GFLOPs		14.1	14.3	14.3	14.3	14.5

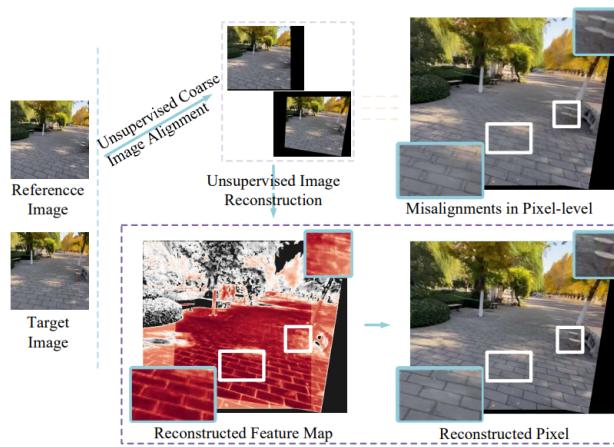
Hình 16: Phân tích so sánh các phương pháp

4.2.2 Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images

- **Tác giả:** Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, Yao Zhao
- **Công bố tại:** IEEE Transactions on Image Processing 2021

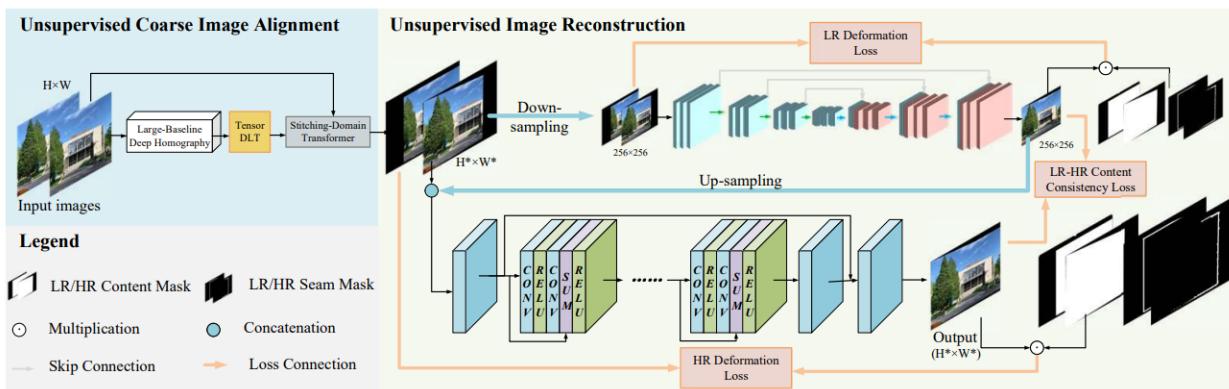
Bài báo này đề xuất một phương pháp ghép ảnh sâu không giám sát mới, giải quyết các hạn chế của các phương pháp truyền thống dựa trên đặc trưng (feature-based) và các phương pháp học sâu có giám sát (supervised deep learning) hiện nay. Phương pháp đề xuất bao gồm hai giai đoạn chính:

1. Căn chỉnh ảnh thô không giám sát (coarse image alignment)
2. Tái thiết ảnh không giám sát (image reconstruction).



Hình 17: Quy trình ghép ảnh sâu không giám sát.

1. Kiến trúc mô hình Mô hình bao gồm hai giai đoạn chính:



Hình 18: Tổng quan về phương pháp học sâu không giám sát

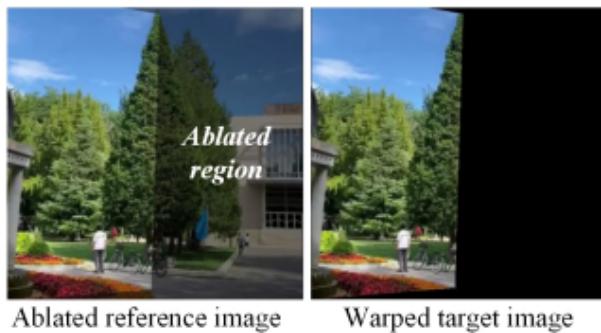
Giai đoạn 1: Unsupervised Coarse Image Alignment

• Mạng ước tính Homography không giám sát:

- Sử dụng kiến trúc mạng đa tỷ lệ (multi-scale) để ước tính ma trận homography giữa hai ảnh đầu vào.
- **Điểm mới:** Thay thế hàm mất mát dựa trên padding (padding-based loss) bằng hàm mất mát dựa trên cắt bỏ (ablation-based loss) để phù hợp với các cảnh có baseline lớn (large-baseline).



Hình 19: Trường hợp thất bại của padding-based

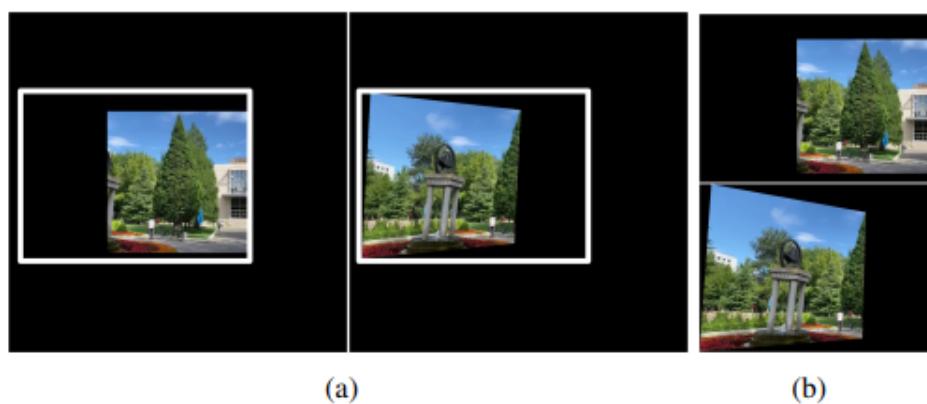


- Hàm mất mát:

$$L'_{PW} = \|\mathcal{H}(E) \odot I^A - \mathcal{H}(I^B)\|_1$$

- **Lớp biến đổi miền ghép ảnh (Stitching-Domain Transformer Layer):**

- **Điểm mới:** Thay thế lớp biến đổi không gian thông thường bằng lớp biến đổi miền ghép ảnh, giúp tiết kiệm không gian bộ nhớ và xử lý ảnh có độ phân giải cao hơn.



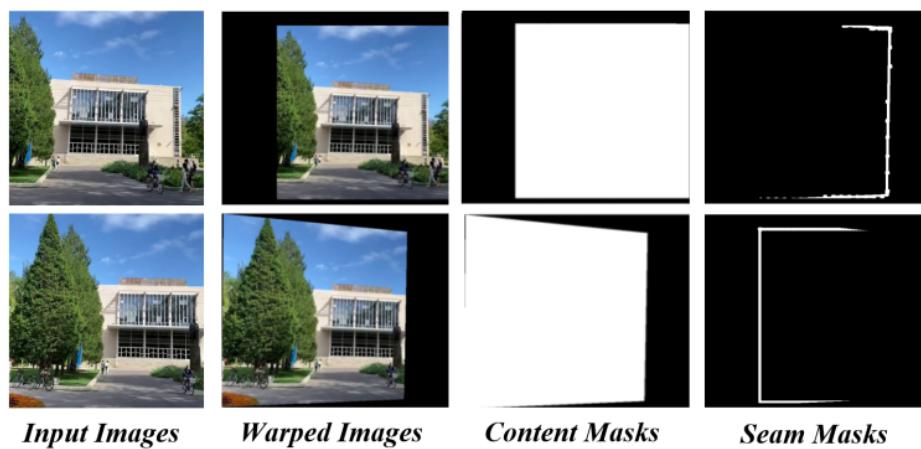
Hình 21: So sánh giữa lớp biến đổi không gian trong phương pháp học sâu đã có so với phương pháp unsupervised trong bài báo (a): Warping của phương pháp đã có (b): Warping của bài báo.

- Biến đổi hai ảnh đầu vào (I_A , I_B) sang miền ghép ảnh (IAW , IBW) dựa trên ma trận homography đã ước tính.

Giai đoạn 2: Unsupervised Image Reconstruction

- Mạng tái thiết ảnh không giám sát:

- Kiến trúc: Bao gồm hai nhánh:
 - Nhánh biến dạng độ phân giải thấp (Low-Resolution Deformation Branch): Mạng encoder-decoder với 3 lớp pooling và 3 lớp deconvolution, kết hợp với skip connection
 - Nhánh tinh chỉnh độ phân giải cao (High-Resolution Refined Branch): Bao gồm các lớp tích chập và resblock để tinh chỉnh chi tiết và nâng cao độ phân giải.
- Mục tiêu: Tái thiết ảnh đã ghép từ đặc trưng sang pixel, loại bỏ các sai lệch (artifacts) do ước tính homography không hoàn hảo.
- Điểm mới: Sử dụng các content masks và seam masks để học quy tắc biến dạng trong quá trình ghép ảnh ở nhánh độ phân giải thấp.



Hình 22: Học các quy tắc biến dạng với mask ở độ phân giải thấp

- Hàm mất mát:

- Content Loss (L'Content): Đảm bảo các đặc trưng của ảnh đã ghép gần giống với các đặc trưng của ảnh đầu vào đã biến đổi.

$$\mathcal{L}_{Content}^l = \mathcal{L}_P(S_{LR} \odot M^{AC}, I^{AW}) + \mathcal{L}_P(S_{LR} \odot M^{BC}, I^{BW})$$

Trong đó:

- $\mathcal{L}_{Content}^l$: Content Loss ở nhánh độ phân giải thấp (low-resolution branch).
- \mathcal{L}_P : Perceptual Loss, sử dụng layer 'conv5_3' của VGG-19.
- S_{LR} : Ảnh đã ghép ở độ phân giải thấp (đầu ra của nhánh độ phân giải thấp).
- M^{AC}, M^{BC} : Content Masks, được tính từ $E_{H \times W}$ (ma trận toàn 1) và I^A, I^B qua công thức (5).
- I^{AW}, I^{BW} : Ảnh đầu vào đã được biến đổi sang miền ghép ảnh (warped images).
- \odot : Phép nhân element-wise (Hadamard product).

- * Seam Loss (L'Seam): Đảm bảo sự chuyển tiếp mượt mà tại các đường nối (seam) giữa hai ảnh.

$$\mathcal{L}_{Seam}^l = \mathcal{L}_1(S_{LR} \odot M^{AS}, I^{AW} \odot M^{AS}) + \mathcal{L}_1(S_{LR} \odot M^{BS}, I^{BW} \odot M^{BS})$$

Trong đó:

- \mathcal{L}_{Seam}^l : Seam Loss ở nhánh độ phân giải thấp.
- \mathcal{L}_1 : L1 Loss.
- M^{AS}, M^{BS} : Seam Masks, được tính từ M^{AC}, M^{BC} qua công thức (6) và (7).
- $I^{AW} \odot M^{AS}$: Phần edge của ảnh warped image A.
- $I^{BW} \odot M^{BS}$: Phần edge của ảnh warped image B.
- * Content Consistency Loss (LCS): Giữ sự nhất quán về nội dung giữa đầu ra của nhánh độ phân giải thấp và nhánh độ phân giải cao.

$$\mathcal{L}_{CS} = \|S_{HR}^{256 \times 256} - S_{LR}\|_1$$

Trong đó:

- \mathcal{L}_{CS} : Content Consistency Loss.
- $S_{HR}^{256 \times 256}$: Ảnh S_{HR} (đầu ra của nhánh độ phân giải cao) được resize về kích thước 256×256 .
- S_{LR} : Ảnh đã ghép ở độ phân giải thấp (đầu ra của nhánh độ phân giải thấp).
- $\|\cdot\|_1$: L1 norm.

2. Quy trình hoạt động

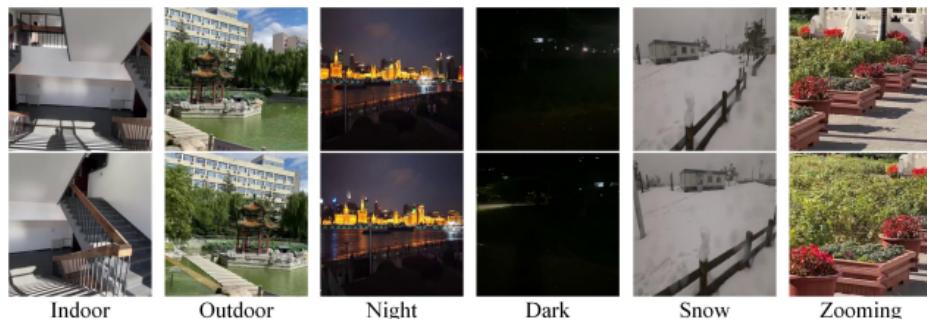
- Căn chỉnh thô:** Hai ảnh đầu vào (IA, IB) được đưa vào mạng ước tính homography để ước tính ma trận homography H.
- Biến đổi:** Hai ảnh đầu vào được biến đổi sang miền ghép ảnh (IAW, IBW) bằng lớp biến đổi miền ghép ảnh.
- Tái thiết ở độ phân giải thấp:** IAW và IBW được down-sample xuống độ phân giải thấp (256×256) và đưa vào nhánh biến dạng độ phân giải thấp để tạo ra ảnh đã ghép SLR ở độ phân giải thấp.
- Tinh chỉnh ở độ phân giải cao:** SLR được up-sample và kết hợp với IAW, IBW để đưa vào nhánh tinh chỉnh độ phân giải cao, tạo ra ảnh đã ghép SHR ở độ phân giải cao.

3. Tập dữ liệu:

- **Train:** Datasets như Microsoft COCO và ImageNet được sử dụng để huấn luyện mô hình với các ảnh đa dạng.
- **Test:** Các bộ dữ liệu được sử dụng để kiểm tra như KITTI, tập trung vào các ảnh ghép nối trong môi trường thực tế với nhiều cảnh vật khác nhau và điều kiện ánh sáng thay đổi.

4. Quá trình huấn luyện:

- **Huấn luyện không giám sát:** Mô hình được huấn luyện hoàn toàn không giám sát, không cần ground truth (ảnh đã ghép hoàn hảo).
- **Dữ liệu huấn luyện:** Tác giả xây dựng một bộ dữ liệu ghép ảnh thực tế lớn, đa dạng về độ phủ (overlap rate), thị sai (parallax), và các điều kiện chụp ảnh (trong nhà, ngoài trời, ban đêm, tuyết, v.v.).



(a) Varying scenes in our dataset.



(b) Varying overlap rates in our dataset.



(c) Varying degrees of parallax in our dataset.

Hình 23: Ảnh minh họa các dataset

• Các bước huấn luyện:

1. Huấn luyện mạng ước tính homography trên tập dữ liệu tổng hợp Stitched MS-COCO [35] (150 epochs).
2. Tinh chỉnh (finetune) mạng ước tính homography trên bộ dữ liệu thực tế (50 epochs).

3. Huấn luyện mạng tái thiết ảnh trên bộ dữ liệu thực tế (20 epochs).

- **Thông số:**

1. Optimizer: Adam [48]
2. Learning rate: 10-4 (giảm dần theo cấp số nhân)
3. Hệ số của các thành phần trong hàm mất mát: As = 2, Ae = 10-6, WLR = 100, WHR = 1, WCS = 1.

5. Đánh giá và so sánh:

Hình 23: So sánh hiệu suất ước tính homography với các phương pháp khác (bao gồm cả có giám sát và không giám sát) trên tập dữ liệu tổng hợp và tập dữ liệu thực tế. Kết quả cho thấy phương pháp đề xuất (Ours-v2) vượt trội hơn các phương pháp khác trên tập dữ liệu thực tế.

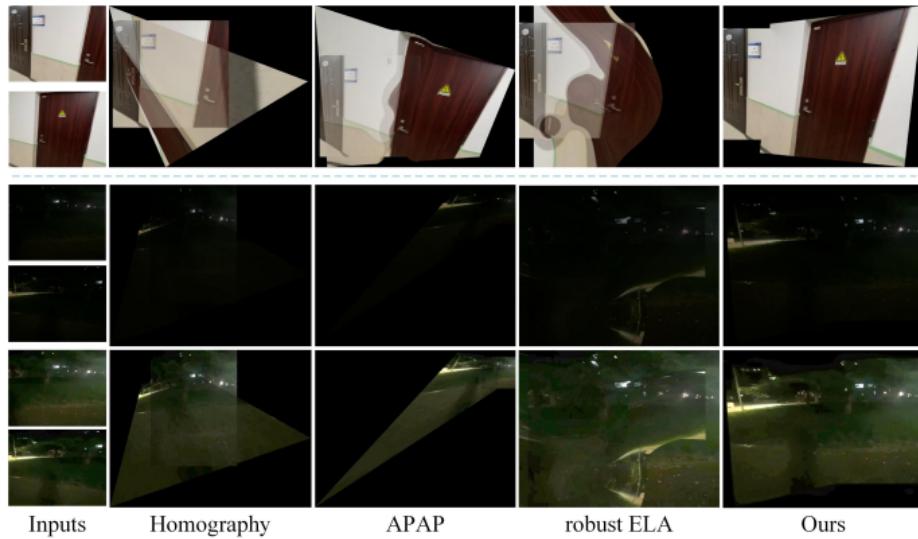
Method	Traditional homography		Deep homography (supervised)		Deep homography (unsupervised)		
	$I_{3 \times 3}$	SIFT [49]+RANSAC [50]	DHN [42]	LB-DHN [36]	UDHN [37]	CA-UDHN [38]	Ours_v1 (synthetic)
Top 0~30%	15.0154	0.6743	3.2998	0.2719	2.1894	15.0082	1.1773
30~60%	18.2515	1.0964	4.8839	0.4140	3.5272	18.2498	1.4544
60~100%	21.3517	19.0286	7.6944	0.9632	6.4984	21.3618	3.0702
Average	18.5220	9.4782	5.5358	0.5962	4.3179	18.5234	2.0239

Method	Traditional homography		Deep homography (supervised)		Deep homography (unsupervised)		
	$I_{3 \times 3}$	SIFT [49]+RANSAC [50]	DHN [42]	LB-DHN [36]	UDHN [37]	Ours_v1 (synthetic)	Ours_v2 (real)
Top 0~30%	16.1923	25.2300	16.3957	24.7515	19.3851	26.1958	27.8386
30~60%	13.0546	22.2308	13.3648	21.1436	15.9251	22.6115	23.9451
60~100%	10.8747	17.5791	11.5001	18.4594	13.1016	19.5277	20.7013
Average	13.1151	21.2541	13.5191	21.1418	15.8252	22.4421	23.8045

Method	Traditional homography		Deep homography (supervised)		Deep homography (unsupervised)		
	$I_{3 \times 3}$	SIFT [49]+RANSAC [50]	DHN [42]	LB-DHN [36]	UDHN [37]	Ours_v1 (synthetic)	Ours_v2 (real)
Top 0~30%	0.3869	0.8598	0.4088	0.8249	0.5732	0.8671	0.9023
30~60%	0.1730	0.7662	0.1699	0.7124	0.3344	0.7844	0.8298
60~100%	0.0732	0.5583	0.0772	0.5497	0.1651	0.6270	0.6846
Average	0.1969	0.7105	0.2042	0.6805	0.3379	0.7456	0.7929

Hình 24: So sánh về ước tính Homography, giải pháp tốt nhất màu đỏ, tốt thứ 2 màu xanh.

Hình 24: So sánh trực quan với các phương pháp ghép ảnh truyền thống trong các trường hợp khó (trong nhà và thiếu sáng).



Hình 25: Các mẫu thử thách để so sánh độ bền một cách trực quan hơn với cảnh trong nhà và bóng tối. Hàng 1: trong nhà. Hàng 2: tối. Hàng 3: hình ảnh tăng cường cho cảnh tối. Độ phân giải của đầu vào là 512×512 .

Hình 25: So sánh về tính liên tục của các đường biên (edge continuity) và sự khác biệt về độ sáng (illumination difference) với phương pháp học sâu có giám sát EPISNet.



Hình 26: Học về tính liên tục và độ chiếu sáng

4.2.3 Implicit Neural Image Stitching With Enhanced and Blended Feature Reconstruction

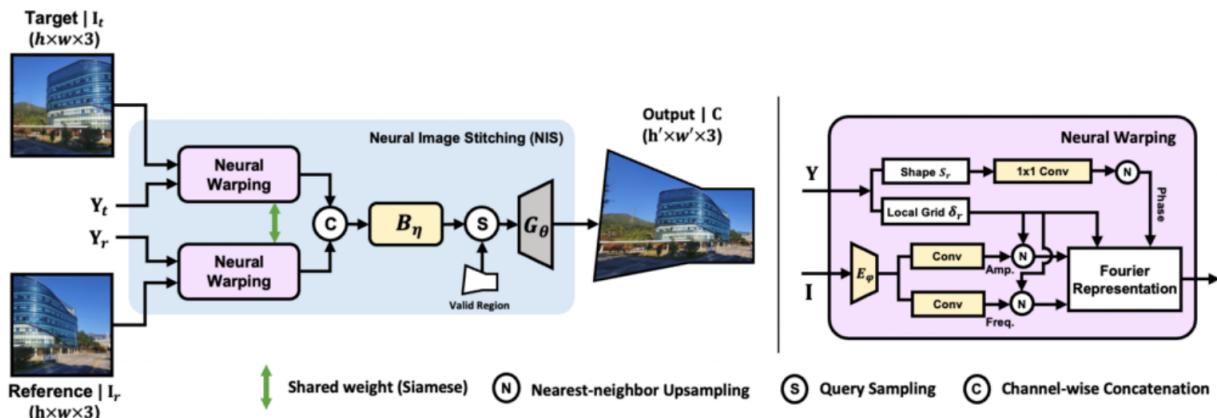
- **Tác giả:** Minsu Kim, Jaewon Lee, Byeonghun Lee, Sunghoon Im, Kyong Hwan Jin

- Công bố tại: WACV 2024

1. Xây dựng bài toán

- **Homography Estimation** Homography là phép biến đổi để căn chỉnh hai hình ảnh qua một lối. Các phương pháp truyền thống như DLT (Direct Linear Transformation) thường dựa trên việc tìm đặc trưng (features) giữa hai ảnh. Tuy nhiên, trong môi trường phức tạp như ngày-kém hoặc với vật thể động, các phương pháp này dễ thất bại. Các bộ ước lượng homography sâu, như IHN, cải thiện khả năng ước lượng bằng cách học đặc trưng qua mạng nơ-ron (CNN). Ví dụ, IHN ước lượng displacement vectors (vectơ dịch chuyển) lặp đi lặp lại qua nhiều bước để hội tụ dần đến phép biến đổi chính xác.
- **Implicit Neural Representation (INR)** INR sử dụng mạng nơ-ron để đại diện tín hiệu liên tục (như ảnh 2D) thông qua hàm ánh xạ. Các ứng dụng bao gồm siêu phân giải (arbitrary-scale SR), tổng hợp cảnh quan, hoặc hình học 3D. Ví dụ, LTE (Local Texture Estimator) kết hợp đặc trưng CNN với tọa độ tương đối, tăng cường tính tổng quát hóa (generalization).
- **Image Blending** Mục tiêu của blending là pha trộn các vùng chồng lấn (overlapping regions) sao cho liền mạch. Các phương pháp từ Poisson blending đến deep blending đều nhắm đến việc xử lý các lỗi ánh sáng và parallax.

2. Kiến trúc mô hình



Hình 27: Kiến trúc mô hình NIS

Phương pháp NIS được mô hình hóa thông qua một hàm ánh xạ:

$$N_\Theta : (y_r, y_t, I_r[x_r], I_t[x_t]) \rightarrow (R, G, B)$$

Với:

- y_r, y_t : tọa độ ảnh tham chiếu và ảnh mục tiêu.
- I_r, I_t : pixel tương ứng từ ảnh tham chiếu và ảnh mục tiêu.
- (R, G, B) : giá trị màu sắc tại điểm được suy diễn.

Hàm N_Θ được chia thành ba thành phần chính:

- **Neural Warping (g):** Học đặc trưng chi tiết tần số cao từ ảnh input.
- **Blender (B_η):** Trộn các đặc trưng đã căn chỉnh từ g .
- **Decoder (G_θ):** Dự đoán RGB từ đặc trưng trộn.

Homography Estimation Homography được ước lượng qua:

$$\hat{D} = \arg \min_D \sum_{k=1}^K \alpha^{K-k} \cdot \|I_r - W(I_t; H_k)\|_1$$

Trong đó:

- $H_k = f_t(D_k, c)$: homography tại bước k từ displacement vector D_k .
- W : phép biến đổi warp.
- K : số lần lặp.
- α : trọng số giảm dần ($\alpha = 0.85$).

3. Phương pháp

Neural Warping (g) Neural warping ước lượng đặc trưng chi tiết cao thông qua:

$$z[y] = g(E_\phi(I_{\text{IN}})[x], c_m)$$

Với:

- E_ϕ : Encoder CNN chuyển đổi ảnh input thành đặc trưng.
- $c_m = y - y'$: tọa độ tương đối.
- $z[y]$: đặc trưng tại điểm y .

Đặc trưng Fourier được sử dụng để cải thiện chi tiết:

$$z[y] = A[y] \cdot (\cos F + \sin F)$$

Trong đó:

- A : Biên độ (amplitude).
- F : Tần số (frequency).
- P : Pha (phase).

Blender (B_η) Blender trộn hai đặc trưng mục tiêu z_t và tham chiếu z_r :

$$C'(y) = B_\eta(z_t, z_r)$$

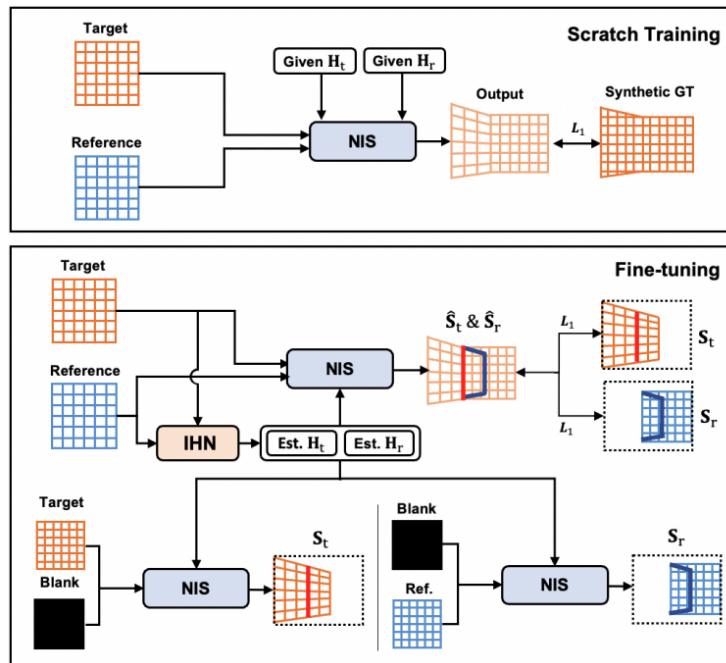
Blender sử dụng một lớp CNN để tạo không gian đặc trưng liền mạch.

Decoder (G_θ) Decoder chuyển không gian đặc trưng sang RGB:

$$\hat{C}[y] = G_\theta(C', y)$$

Với y là tọa độ pixel.

4. Chi tiết chiến lược đào tạo mô hình



Hình 28: Chiến lược đào tạo

NIS được huấn luyện theo chiến lược hai giai đoạn để tập trung vào các nhiệm vụ riêng biệt: tăng cường chi tiết ảnh và pha trộn đặc trưng.

4.1 Giai đoạn 1: Học chi tiết tần số cao Mục tiêu: Mô hình học cách tái tạo chi tiết tần số cao thông qua dữ liệu tổng hợp, tức là ảnh đã được biến dạng với phép biến đổi homography được biết trước.

Phương pháp:

- Ảnh tổng hợp được tạo bằng cách áp dụng homography ngẫu nhiên lên các ảnh gốc từ tập MS-COCO.
- Mô hình được tối ưu hóa với hàm mất mát L_1 :

$$L_1(\Theta) = \arg \min_{\Theta} \sum_{y_u} \|C[y_u] - \hat{C}[y_u]\|_1$$

Trong đó:

- C : Ảnh mục tiêu.
- \hat{C} : Ảnh dự đoán từ mô hình tại tọa độ y_u .

Tăng cường dữ liệu:

- Tạo các biến dạng bằng cách dịch chuyển góc ảnh trong khoảng 25
- Phép biến đổi homography được thực hiện trên từng minibatch.

Kết quả: Mô hình học cách tái tạo ảnh có chi tiết tần số cao từ đặc trưng Fourier, cải thiện độ sắc nét và giảm mờ.

4.2 Giai đoạn 2: Học pha trộn đặc trưng Mục tiêu: Giải quyết sự không khớp màu sắc (color mismatch) và lỗi parallax trong các vùng chồng lấn của ảnh.

Phương pháp:

- Mô hình sử dụng IHN để ước lượng phép biến đổi homography trên các ảnh thực (UDIS-D), do dữ liệu thực không có ground truth.
- Hàm mất mát seam loss được áp dụng để điều chỉnh vùng giao nhau (seam region):

$$L_{\text{seam}}(\Theta) = \arg \min_{\Theta} \sum_n \|\hat{S}_n - S_n\|_1$$

Trong đó:

- S_n : Đặc trưng tham chiếu tại vùng giao nhau.
- \hat{S}_n : Đặc trưng được pha trộn bởi Blender (B_η).
- M_n : Mặt nạ (mask) để xác định vùng seam.

Chi tiết triển khai:

- Sử dụng ảnh trống (blank images) để tạo các tham chiếu nội bộ nhằm khắc phục sự thiếu ground truth trên vùng giao nhau.
- Đóng bằng các phần mô hình đã được học ở Giai đoạn 1, chỉ tối ưu hóa Blender và Decoder.

4.3 Chi tiết kỹ thuật huấn luyện

Giai đoạn 1:

- Batch size: 20.
- Số vòng lặp: 250,000.
- Crop size: 48×48 .
- Optimizer: Adam ($\beta_1 = 0.9, \beta_2 = 0.999$).
- Learning rate: 10^{-4} , giảm dần mỗi epoch theo hàm mũ (0.98).

Giai đoạn 2:

- Batch size: 1.

- Số vòng lặp: 60,000 đến 300,000 (tùy thuộc dữ liệu).
- Kích thước ảnh: 128×128 .
- Learning rate: tương tự Giai đoạn 1.

5. Thử nghiệm

5.1 Bộ dữ liệu

- **MS-COCO:** Dùng để tổng hợp ảnh với phép biến dạng đồng nhất, không có lỗi parallax.
- **UDIS-D:** Dữ liệu thực với nhiều lỗi parallax.

5.2 Chi tiết cài đặt

- Sử dụng IHN để căn chỉnh ảnh, sau đó huấn luyện NIS với grids đã căn chỉnh.
- Kiến trúc mạng:
 - Encoder và Blender sử dụng EDSR (Enhanced Deep Residual Networks).
 - Decoder là một MLP 4 lớp với 256 đơn vị ẩn.

5.3 Đánh giá

Method	mPSNR (\uparrow)	mSSIM (\uparrow)	#Params.
Bilinear	34.78	0.96	-
Bicubic	36.25	0.97	-
UDIS [33]	33.45	0.97	8.0M
NIS (<i>ours</i>)	38.69	0.98	3.2M

(a) Evaluation on Synthetic Images.

Method	NIQE (\downarrow)	PIQE (\downarrow)	BRISQUE (\downarrow)
APAP [48]	3.30	46.95	34.72
Robust ELA [21]	3.59	53.67	37.78
SPW [23]	3.39	51.68	36.84
LPC [13]	3.37	50.81	37.15
LPC + Graph Cut	3.50	50.63	37.14
UDIS [33]	3.43	50.01	36.71
IHN [4]+NIS	3.28	46.21	33.17
IHN [4]+NIS (F)	3.15	43.05	31.14

Hình 29: So sánh định lượng về hiệu suất khâu. Màu đỏ và màu xanh lam lần lượt biểu thị hiệu suất tốt nhất và hiệu suất tốt thứ hai.

- **Kết quả định lượng:** Trên MS-COCO, NIS đạt mPSNR cao hơn UDIS (38.69 so với 33.45) và giảm lỗi màu sắc.

- Kết quả định tính:** Các hình ảnh từ NIS có chi tiết rõ ràng và ít lỗi ánh sáng/parallax hơn so với các phương pháp khác.
- 6. Nghiên cứu cắt bỏ:** Loại bỏ từng thành phần Fourier (biên độ, tần số, pha) cho thấy tầm quan trọng của từng yếu tố. Chiến lược học hai giai đoạn giúp cải thiện cả chi tiết và độ liền mạch ảnh.

4.2.4 Bảng so sánh các phương pháp học sâu

Tiêu chí	Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images	Implicit Neural Image Stitching With Enhanced and Blended Feature Reconstruction	Efficient Unsupervised Image Stitching Using Attention Mechanism with Deep Homography Estimation
Phương pháp	Sử dụng học không giám sát kết hợp mạng CNN để học các đặc trưng từ các ảnh, sau đó tái tạo lại một bức ảnh ghép mới	Dự đoán các đặc trưng Fourier của hai ảnh đầu vào, kết hợp thông qua module blender để tạo tín hiệu gộp liền mạch, sử dụng MLP decoder để giải mã tín hiệu gộp và xuất ra ảnh ghép hoàn chỉnh.	Lấy YOLOv8 làm nền tảng, tinh chỉ, áp dụng cơ chế chú ý và module pooling kim tự tháp để trích đặc trưng; tích hợp mạng homography để ước tính phép biến đổi hình học, Các phép ghép sơ bộ và căn chỉnh được thực hiện qua các mặt nạ (masks) nhằm giảm lỗi ghép.
Đánh giá	KITTI: SSIM: 0.7929 ; PSNR: 23.8045	MS-COCO: mSSIM: 0.98 ; mPSNR: 38.69	UDIS: SSIM: 0.8414 ; PSNR: 26.34 MSV-Synth: SSIM: 0.8494 ; PSNR: 26.42

Hình 30: Bảng so sánh 1

Tiêu chí	Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images	Implicit Neural Image Stitching With Enhanced and Blended Feature Reconstruction	Efficient Unsupervised Image Stitching Using Attention Mechanism with Deep Homography Estimation
Đầu vào	Tập ảnh có sự chồng lấn, tiền xử lý chuẩn hóa về cùng kích thước ảnh	Tập ảnh có sự chồng lấn Một lưới biến đổi (transformation grid)	Tập ảnh có sự chồng lấn đã được tiền xử lý về kích thước 640 x 640
Đầu ra	Một bức ảnh ghép hoàn chỉnh	Một bức ảnh ghép hoàn chỉnh	Một bức ảnh ghép hoàn chỉnh
Tập dữ liệu	Training: Microsoft COCO, ImageNet Testing: KITTI	Training: MS-COCO, UDIS-D Testing: UDIS-D, SUN360	Pre-trained: Warped COCO Training: UDIS, MVS-Synth Testing: UDIS, MVS-Synth

Hình 31: Bảng so sánh 2

5 Phương pháp

5.1 Bộ dữ liệu huấn luyện

UDIS-D là một bộ dữ liệu lớn được thiết kế cho bài toán ghép ảnh (image stitching) trong môi trường không giám sát. Bộ dữ liệu này được giới thiệu trong bài báo *“Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images”* của Nie và cộng sự. Bộ dữ liệu được công bố tại: **UDIS-D**.

Bộ dữ liệu được xây dựng từ video chuyển động thay đổi, với nguồn gốc bao gồm:

- Một phần từ tài nguyên trong bài báo *Content-aware unsupervised deep homography estimation*.
- Một phần do nhóm tác giả tự xây dựng bằng cách trích xuất các khung hình từ các video với khoảng thời gian khác nhau.

Tính đa dạng của UDIS-D:

- **Cảnh quan:** Bộ dữ liệu bao gồm các khung cảnh thực tế đa dạng như trong nhà, ngoài trời, ban đêm, ánh sáng yếu, môi trường có tuyết hoặc các cảnh phóng to.



Hình 32: Đa dạng về cảnh quan

- **Tỷ lệ chồng lấp (overlap rates):** Các mẫu có tỷ lệ chồng lấp khác nhau với ba mức độ:
 - Cao: Lớn hơn 90%.
 - Trung bình: 60%-90%.
 - Thấp: Dưới 60%.

Trung bình, tỷ lệ chồng lấp của dataset lớn hơn 90%.



High overlap rate

Middle overlap rate

Low overlap rate

Hình 33: Da dạng về tỷ lệ chồng lấp

- **Mức độ thị sai (parallax):** Mô phỏng thị sai thực tế với:

- Thị sai nhỏ: Sai số nhỏ hơn 30 pixel (91% mẫu).
- Thị sai lớn: Sai số lớn hơn 30 pixel (9% mẫu).



Small parallax

Large parallax

Hình 34: Da dạng về mức độ thị sai

Thống kê dữ liệu:

- Số lượng mẫu:
 - 10,440 mẫu dùng để huấn luyện.
 - 1,106 mẫu dùng để kiểm tra.
- Phân bố dữ liệu:
 - Tỷ lệ chồng lấp:
 - * Cao: 16%.
 - * Trung bình: 66%.
 - * Thấp: 18%.
 - Mức độ thị sai:

- * Thị sai nhỏ: 91%.
- * Thị sai lớn: 9%.

Phương pháp đo lường và sử dụng: Bộ dữ liệu không chứa ground-truth, nhưng đi kèm kết quả kiểm tra của nhóm tác giả được tạo ra từ mô hình của bài báo *Ünsupervised Deep Image Stitching: Reconstructing Stitched Features to Images*. Để đo lường thị sai, hình ảnh mục tiêu được căn chỉnh với hình ảnh tham chiếu bằng một phép đồng nhất toàn cục, sau đó tính sai số tối đa giữa các điểm đặc trưng trong hình ảnh đã căn chỉnh.

5.2 Phương pháp truyền thống

Nhóm đã lựa chọn triển khai và đánh giá kỹ thuật ghép nối đặc trưng **SIFT (Scale-Invariant Feature Transform)** trong giai đoạn đăng ký hình ảnh, bởi đây là một trong những phương pháp phổ biến và được đánh giá cao về tính ổn định trong các nghiên cứu trước đó. SIFT nổi bật với khả năng nhận diện và so khớp đặc trưng bất biến trước các biến dạng như xoay, tỷ lệ, và thay đổi góc nhìn, điều này đặc biệt quan trọng đối với các bài toán ghép ảnh đòi hỏi độ chính xác cao.

1. Phát hiện và mô tả các điểm đặc trưng (Keypoint Detection & Descriptor)

1.1 Điều kiện phát hiện keypoint (Interest Point):

Các điểm đặc trưng trong ảnh được phát hiện thông qua việc sử dụng **máy lọc DoG (Difference of Gaussian)**. Quá trình này tạo ra các cấp độ khác nhau của ảnh gốc bằng cách áp dụng một hàm lọc Gaussian với các kích cỡ khác nhau.

Công thức tính **Ảnh DoG (Difference of Gaussian)**:

$$DoG(x, y, \sigma) = G(x, y, k\sigma) - G(x, y, \sigma)$$

Trong đó:

- $G(x, y, \sigma)$ là hàm Gaussian với độ rộng chuẩn σ ,
- k là một hằng số tỷ lệ dùng để điều chỉnh độ rộng chuẩn.

1.2 Đặc trưng hình ảnh (Descriptor):

Đặc trưng của ảnh tại mỗi điểm góc được tính toán thông qua việc tính các vector gradient tại các điểm này. Cụ thể, độ lớn và góc gradient được tính theo công thức dưới đây:

$$\text{Gradient Magnitude: } G = \sqrt{(I_x)^2 + (I_y)^2}$$

$$\text{Gradient Angle: } \theta = \tan^{-1} \left(\frac{I_y}{I_x} \right)$$

Trong đó:

- I_x và I_y lần lượt là các đạo hàm của ảnh theo hướng ngang và dọc,
- G là độ lớn gradient,

- θ là góc gradient tại mỗi điểm đặc trưng.

2. Khớp đặc trưng (Keypoint Matching)

Quá trình khớp đặc trưng giữa hai ảnh được thực hiện bằng cách đo khoảng cách giữa các vector đặc trưng của chúng. Phương pháp phổ biến nhất là sử dụng **L2 distance (Euclidean distance)** để tính toán khoảng cách giữa các vector mô tả đặc trưng.

Công thức tính khoảng cách L2 giữa hai vector đặc trưng:

$$\text{Distance}(f_1, f_2) = \sqrt{\sum_{i=1}^n (f_{1i} - f_{2i})^2}$$

Trong đó:

- f_1 và f_2 là các vector đặc trưng của hai điểm đặc trưng tương ứng,
- n là số chiều của vector đặc trưng.

Sau khi tính toán khoảng cách giữa tất cả các cặp vector đặc trưng, các điểm khớp sẽ được chọn dựa trên những cặp có khoảng cách nhỏ nhất.

3. Ước lượng Ma trận Homography

Sau khi các điểm đặc trưng đã được khớp, bước tiếp theo là tính toán ma trận Homography. Ma trận Homography là một phép biến đổi hình học giúp chuyển các điểm từ một ảnh này sang ảnh khác. Ma trận này có thể được ước lượng thông qua phương pháp **RANSAC** (Random Sample Consensus), giúp loại bỏ các điểm khớp sai (outliers).

Ma trận Homography H được tính theo công thức sau:

$$p' = H \cdot p$$

Trong đó:

- $p = (x, y)^T$ là tọa độ điểm trên ảnh đầu vào,
- $p' = (x', y')^T$ là tọa độ điểm tương ứng trên ảnh thứ hai,
- H là ma trận 3×3 , đại diện cho phép biến đổi hình học giữa hai ảnh.

Quá trình ước lượng này sử dụng thuật toán RANSAC để tìm bộ điểm khớp đúng bằng cách chọn mẫu và ước tính Homography từ các điểm thử nghiệm, đồng thời loại bỏ các điểm không phù hợp.

4. Chuyển ảnh với Ma trận Homography (Image Warping)

Một bước quan trọng trong quá trình ghép ảnh là chuyển ảnh theo ma trận Homography. Quy trình này sử dụng công thức biến hình học để ánh xạ các điểm từ ảnh đầu tiên sang ảnh thứ hai (hoặc từ ảnh thứ hai sang một ảnh phối hợp).

Công thức biến đổi một điểm $p = (x, y)$ trong không gian ảnh sang một điểm $p' = (x', y')$ là:

$$p' = H \cdot p$$

Quá trình này là cơ sở để thực hiện việc "warp" (biến dạng) các ảnh sao cho chúng có thể được kết hợp với nhau tạo thành một panorama.

5. Tạo Mặt nạ và Hòa trộn (Masking and Blending)

Để đảm bảo rằng việc ghép ảnh diễn ra mượt mà và không có các rìa răng cưa giữa các bức ảnh, ta cần tạo ra mặt nạ *mask*. Mặt nạ này được xây dựng bằng cách sử dụng hàm Gauss để làm mịn vùng giao nhau giữa các ảnh.

Mặt nạ được xác định bằng công thức Gaussian, với độ rộng cửa sổ smoothing σ :

$$\text{Mask}(x) = e^{-\frac{x^2}{2\sigma^2}}$$

Trong đó σ là độ rộng của cửa sổ smoothing, giúp tạo ra sự chuyển tiếp mượt mà giữa các vùng ảnh được ghép.

Khi sử dụng mặt nạ, quá trình hòa trộn ảnh sẽ được tính toán như sau:

$$I_{final}(x, y) = w_1(x, y) \cdot I_1(x, y) + w_2(x, y) \cdot I_2(x, y)$$

Trong đó I_1 và I_2 là các ảnh đã được warp, và w_1, w_2 là các mặt nạ đối ứng, đảm bảo rằng hai ảnh được hòa trộn một cách mượt mà ở vùng giao nhau.

6. Lọc và loại bỏ không gian thừa

Sau khi ghép ảnh, một bước quan trọng là loại bỏ những không gian đen (không có dữ liệu ảnh). Việc này được thực hiện bằng cách xác định các chỉ số hàng và cột mà tại đó các giá trị khác không phải 0 (tức là có dữ liệu ảnh) và cắt bớt không gian thừa bên ngoài.

Công thức để cắt ảnh và giữ lại các vùng có dữ liệu là:

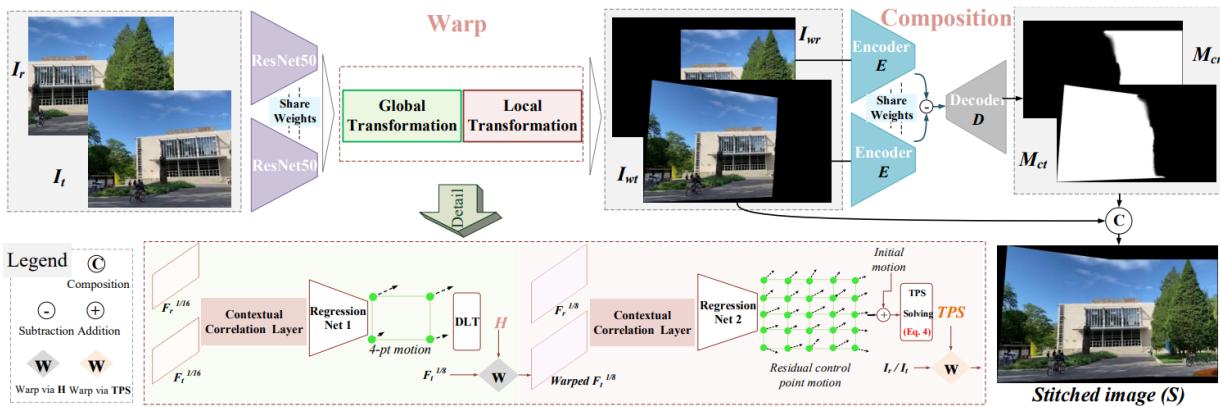
$$I_{final} = I_{final}[min_row : max_row, min_col : max_col]$$

Tại đây, các chỉ số min_row, max_row và min_col, max_col được tính từ các điểm có chứa dữ liệu ảnh.

5.3 Phương pháp học sâu

Nhằm áp dụng mô hình tiên tiến nhất vào bài nghiên khảo sát của nhóm, nhóm đã quyết định triển khai và đánh giá phương pháp được trình bày trong bài báo "**Parallax-Tolerant Unsupervised Deep Image Stitching**" là bài báo được cải tiến từ công trình "**Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images**"

5.3.1 Kiến trúc mô hình



Hình 35: Kiến trúc mô hình

Tổng quan

Hình ảnh tham chiếu I_r và hình ảnh mục tiêu I_t có vùng chồng lấn được nhập vào mạng. Quá trình hoạt động của mô hình như sau:

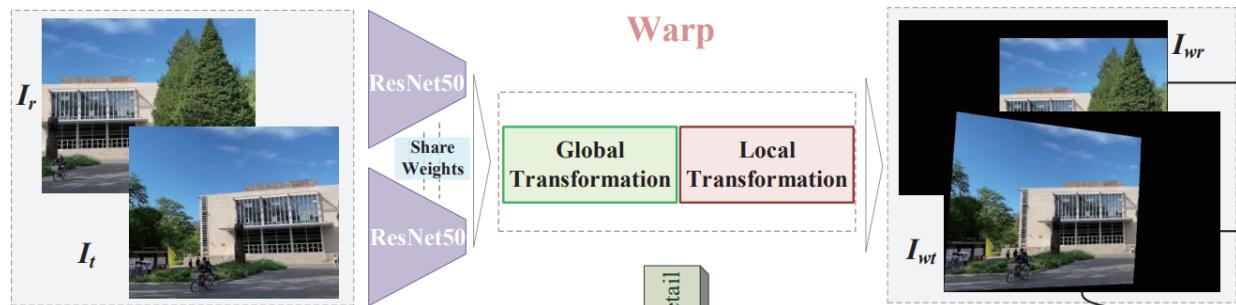
- **Giai đoạn warp:** Tính toán các phép biến dạng toàn cục (homography) và cục bộ (TPS) để căn chỉnh nội dung và bảo toàn hình dạng.
- **Giai đoạn composition:** Sử dụng mang học sâu để tạo ra mặt nạ tổng hợp M_{cr} và M_{ct} giúp loại bỏ các hiện vật (artifacts).

Hình ảnh tổng hợp S được tạo bởi:

$$S = M_{cr} \times I_{wr} + M_{ct} \times I_{wt},$$

với I_{wr} và I_{wt} là các hình ảnh tham chiếu và mục tiêu sau khi được biến dạng.

1. Giai đoạn Warp không giám sát



Hình 36: Kiến trúc giai đoạn Warp

1.1. Tham số hóa Warp

Biến đổi Homography: Là một phép biến đổi tuyến tính toàn cục giữa hai hình ảnh, được biểu diễn qua một ma trận 3×3 . Phép biến đổi này được tham số hóa bằng chuyển

động của bốn đỉnh (4-pt parameterization) và được tính toán bằng phương pháp Direct Linear Transform (DLT).

Biến đổi Thin-Plate Spline (TPS): Là một phép biến đổi phi tuyến. Phép biến đổi được xác định bởi hai tập hợp các điểm điều khiển P trên ảnh phẳng và P' trên ảnh biến dạng:

$$P = [p_1, p_2, \dots, p_N]^T, \quad P' = [p'_1, p'_2, \dots, p'_N]^T,$$

trong đó $p_i, p'_i \in \mathbb{R}^{2 \times 1}$.

Công thức phép biến đổi TPS:

$$p' = T(p) = C + Mp + \sum_{i=1}^N w_i O(\|p - p_i\|^2),$$

với:

$$C \in \mathbb{R}^{2 \times 1}, \quad M \in \mathbb{R}^{2 \times 2}, \quad w_i \in \mathbb{R}^{2 \times 1}$$

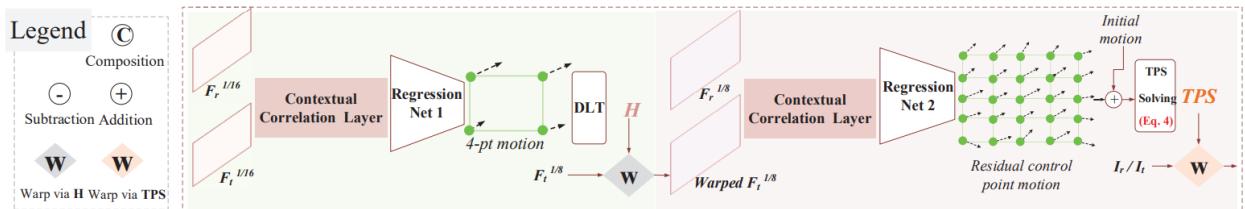
là các tham số của phép biến đổi. Hàm cơ sở hướng tâm (radial basis function) được định nghĩa như sau:

$$O(r) = r^2 \log(r^2).$$

Các ràng buộc để đảm bảo sự liên tục và không bị biến dạng của TPS:

$$\sum_{i=1}^N w_i = 0, \quad \sum_{i=1}^N p_i w_i^T = 0.$$

1.2. Quy trình Warp



Hình 37: Chi tiết giai đoạn Warp

Kiến trúc mạng: Sử dụng ResNet50 để trích xuất đặc trưng từ ảnh tham chiếu I_r và ảnh mục tiêu I_t . Các đặc trưng này được xử lý bởi lớp tương quan ngữ cảnh (contextual correlation layer) để tính toán các chuyển động ban đầu (4-pt motion). Sau đó, chuyển động dư thừa (residual motion) được dự đoán để hiệu chỉnh TPS.

Sử dụng mô hình ResNet50:

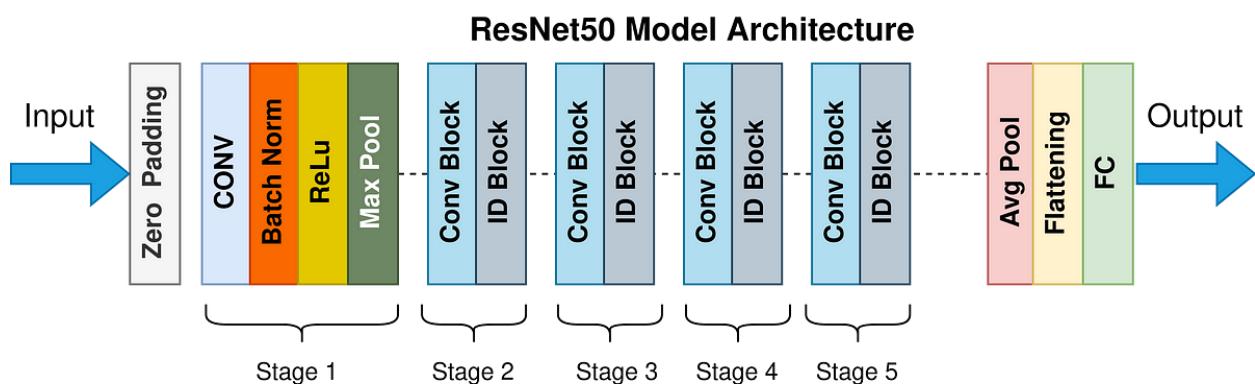
- Mô hình chính cho việc trích xuất đặc trưng.

- Dựa trên các đặc trưng này để tính toán tham số biến dạng (TPS + Homography).

Chi tiết về cách sử dụng mô hình ResNet50 trong trích xuất đặc trưng và tính toán tham số biến dạng:

1. Trích xuất đặc trưng với ResNet50:

- ResNet50 là một mạng nơ-ron sâu bao gồm 50 lớp, sử dụng các *residual blocks* để cải thiện khả năng học và giảm hiện tượng vanishing gradient.
- Khi đưa vào một ảnh, ResNet50 sẽ học được các đặc trưng của ảnh qua các lớp convolutional và activation functions, tạo ra một vector đặc trưng biểu diễn các tính chất của ảnh (chẳng hạn như đường viền, đối tượng, kết cấu).



Hình 38: Kiến trúc ResNet50

2. Sử dụng các đặc trưng để tính toán tham số biến dạng (TPS + Homography):

- Sau khi trích xuất đặc trưng, sử dụng chúng để tính toán các tham số biến dạng như **TPS** (Thin Plate Spline) và **Homography**.

1.3. Tối ưu hóa Warp

Hàm mất mát trong giai đoạn warp gồm hai thành phần:

Alignment Loss $L_{w_alignment}$:

Đảm bảo căn chỉnh chính xác các vùng chồng lấn giữa hai ảnh:

$$\begin{aligned} L_{w_alignment} = & \lambda \|I_r \cdot \varphi(1, H) - \varphi(I_t, H)\|_1 + \\ & \lambda \|I_t \cdot \varphi(1, H^{-1}) - \varphi(I_r, H^{-1})\|_1 + \\ & \|I_r \cdot \varphi(1, TPS) - \varphi(I_t, TPS)\|_1 \end{aligned}$$

với H là tham số homography và TPS là tham số Thin Plate Spline.

Distortion Loss $L_{w_distortion}$:

Giảm biến dạng trong vùng không chồng lấn bằng hai ràng buộc:

Ràng buộc trong lưới Hạn chế méo mó lưới:

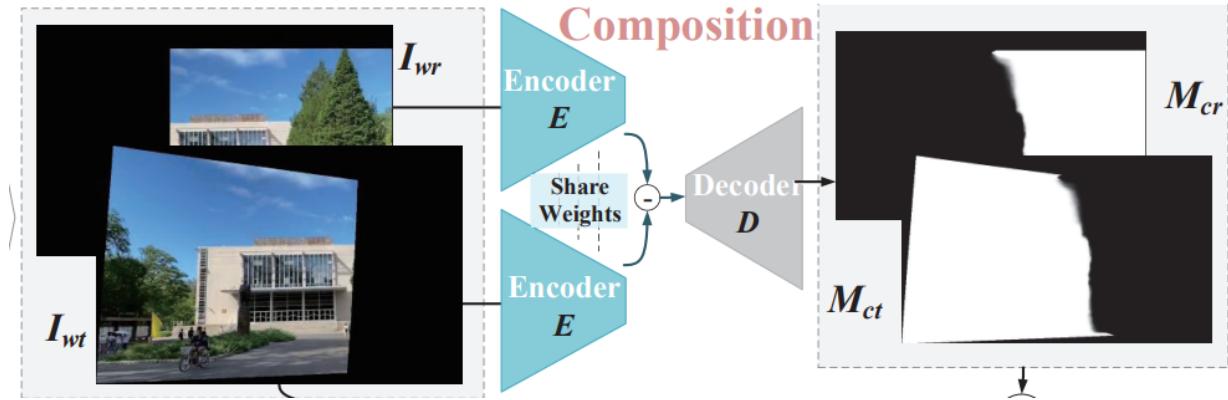
$$\ell_{intra} = \frac{1}{(U+1) \times V} \sum_{e_{hor}} \sigma(\langle e, \hat{i} \rangle - \frac{2W}{V}) + \frac{1}{U \times (V+1)} \sum_{e_{ver}} \sigma(\langle e, \hat{j} \rangle - \frac{2H}{U}),$$

với \hat{i} và \hat{j} là vector đơn vị ngang/dọc.

Ràng buộc giữa các lối Duy trì cấu trúc hình học:

$$\ell_{inter} = \frac{1}{Q} \sum_{\{e_{s1}, e_{s2}\}} S_{s1, s2} \cdot \left(1 - \frac{\langle e_{s1}, e_{s2} \rangle}{\|e_{s1}\| \cdot \|e_{s2}\|} \right).$$

2. Giai đoạn Composition liền mạch không giám sát

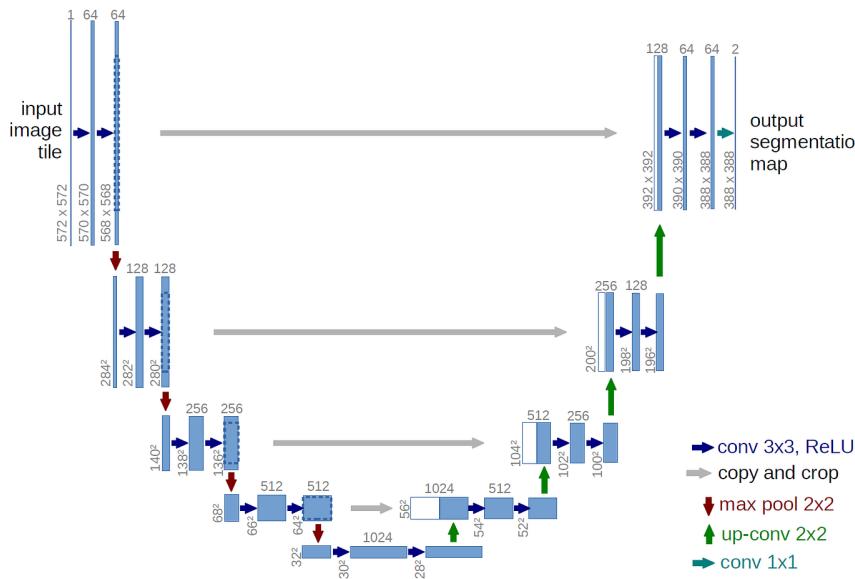


Hình 39: Giai đoạn Composition

Mục đích: Phương pháp truyền thống dựa trên *seam cutting* hoặc tái tạo không xử lý lý tốt các hiện vật parallax. Giai đoạn này đề xuất một cách tiếp cận học sâu không giám sát để học các mặt nạ tổng hợp (composition masks) bằng các ràng buộc được thiết kế đặc biệt.

2.1. Quy trình Composition

Sử dụng mạng UNet để tạo ra mặt nạ M_{cr} và M_{ct} từ các ảnh đã được biến dạng I_{wr} và I_{wt} . Đặc trưng của từng ảnh được xử lý riêng rẽ bằng encoder có trọng số chia sẻ.



Hình 40: Kiến trúc UNet

2.2. Tối ưu hóa Composition

Hàm mất mát của giai đoạn này gồm hai thành phần:

Boundary Term $L_{c_boundary}$:

Ràng buộc các pixel ở ranh giới vùng chồng lấn:

$$L_{c_boundary} = \|(S - I_{wr}) \cdot M_{br}\|_1 + \|(S - I_{wt}) \cdot M_{bt}\|_1.$$

Smoothness Term $L_{c_smoothness}$:

Dảm bảo sự chuyển tiếp mượt mà giữa hai ảnh:

$$L_{c_smoothness} = \ell_D + \ell_S,$$

với ℓ_D là sự mượt mà trên bản đồ chênh lệch và ℓ_S là sự mượt mà trên ảnh tổng hợp.

3. Lặp lại Điều chỉnh Biến dạng

Để cải thiện khả năng tổng quát hóa, tác giả đề xuất chiến lược điều chỉnh lặp lại (iterative adaption). Hàm mất mát trong quá trình này:

$$L_{w_adaption} = \|I_r \cdot \varphi(1, TPS) - \varphi(I_t, TPS)\|_1.$$

5.3.2 Quá trình huấn luyện

Thông tin chi tiết:

- Mạng huấn luyện trên PyTorch với GPU NVIDIA RTX 3090 Ti.
- Sử dụng Adam Optimizer với learning rate ban đầu là 10^{-4} , giảm theo hàm số mũ trong quá trình huấn luyện.

- Quá trình huấn luyện được chia ra làm 2 giai đoạn:
 - Giai đoạn “Warp”: Tạo ra một phép biến dạng chính xác để căn chỉnh các hình ảnh.
 - Giai đoạn “Composition”: Tạo ra hình ảnh tổng hợp liền mạch từ các hình ảnh đã được biến dạng.
- Tổng số epoch:
 - 100 epoch cho giai đoạn “Warp”.
 - 50 epoch cho giai đoạn “Composition”.

Huấn luyện giai đoạn Warp (biến dạng hình ảnh): Mục tiêu là thực hiện đồng nhất nội dung và bảo toàn hình dạng.

Tham số chính:

- $\omega = 10$: Trọng số cho măt măt về biến dạng.
- $\lambda = 3$: Trọng số cân bằng tác động của các phép biến đổi khác nhau.

Tối ưu hàm măt măt: Tổng măt măt cho giai đoạn warp là:

$$L_w = L_w^{\text{alignment}} + \omega L_w^{\text{distortion}}$$

- $L_w^{\text{alignment}}$: Đảm bảo vùng chồng lấn giữa các ảnh được căn chỉnh chính xác ở cấp độ pixel. Sử dụng phép đo khác biệt pixel giữa hai ảnh chồng lấn.
- $L_w^{\text{distortion}}$: Bảo toàn cấu trúc hình học bằng cách hạn chế các méo mó trong vùng không chồng lấn.

Huấn luyện giai đoạn Composition (tổng hợp hình ảnh): Mục tiêu là tạo ra hình ảnh liền mạch bằng cách tìm đường “seam” giữa các hình ảnh chồng lấn.

Tham số chính:

- $\alpha = 10,000, \beta = 1,000$: Cân bằng giữa các yếu tố măt măt trong giai đoạn tổng hợp.

Tối ưu hàm măt măt: Tổng măt măt cho giai đoạn này là:

$$L_c = \alpha L_c^{\text{boundary}} + \beta L_c^{\text{smoothness}}$$

- L_c^{boundary} : Đảm bảo các đường viền “seam” nằm trong vùng giao giữa các hình ảnh.
- $L_c^{\text{smoothness}}$: Hạn chế sự không liên tục của “seam” và tạo sự chuyển tiếp mượt mà trên đường nối.

6 Cài đặt và thử nghiệm

6.1 Cấu hình môi trường

- GPU: NVIDIA GeForce MX450 (10GB)
- CUDA: 12.1
- Python: 3.11.7

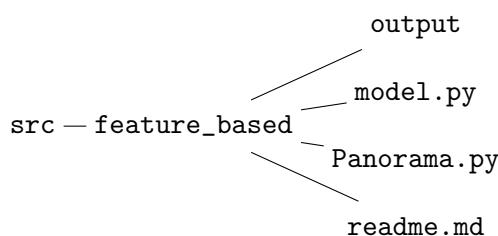
Các thư viện Python sử dụng:

- numpy: 1.26.3
- opencv-python: 4.10.0.84
- scikit-image: 0.25.0
- scipy: 1.14.1
- torch: 2.3.1+cu121
- torchaudio: 2.3.1+cu121
- torchvision: 0.18.1+cu121

Các yêu cầu khác được mô tả chi tiết trong tệp requirements.txt. Truy cập và cài đặt requirements.txt để biết thêm chi tiết.

6.2 Phương pháp truyền thống

Cấu trúc tổ chức mã nguồn phương pháp truyền thống:



Chi tiết các bước kiểm thử:

1. Diều hướng đến thư mục src và sau đó vào thư mục feature_based.
2. Diều chỉnh đường dẫn dữ liệu thử nghiệm trong tệp Panorama.py sao cho phù hợp với cấu trúc thư mục hiện tại, đảm bảo chính xác. Sau đó, thực thi tệp Panorama.py với cú pháp:

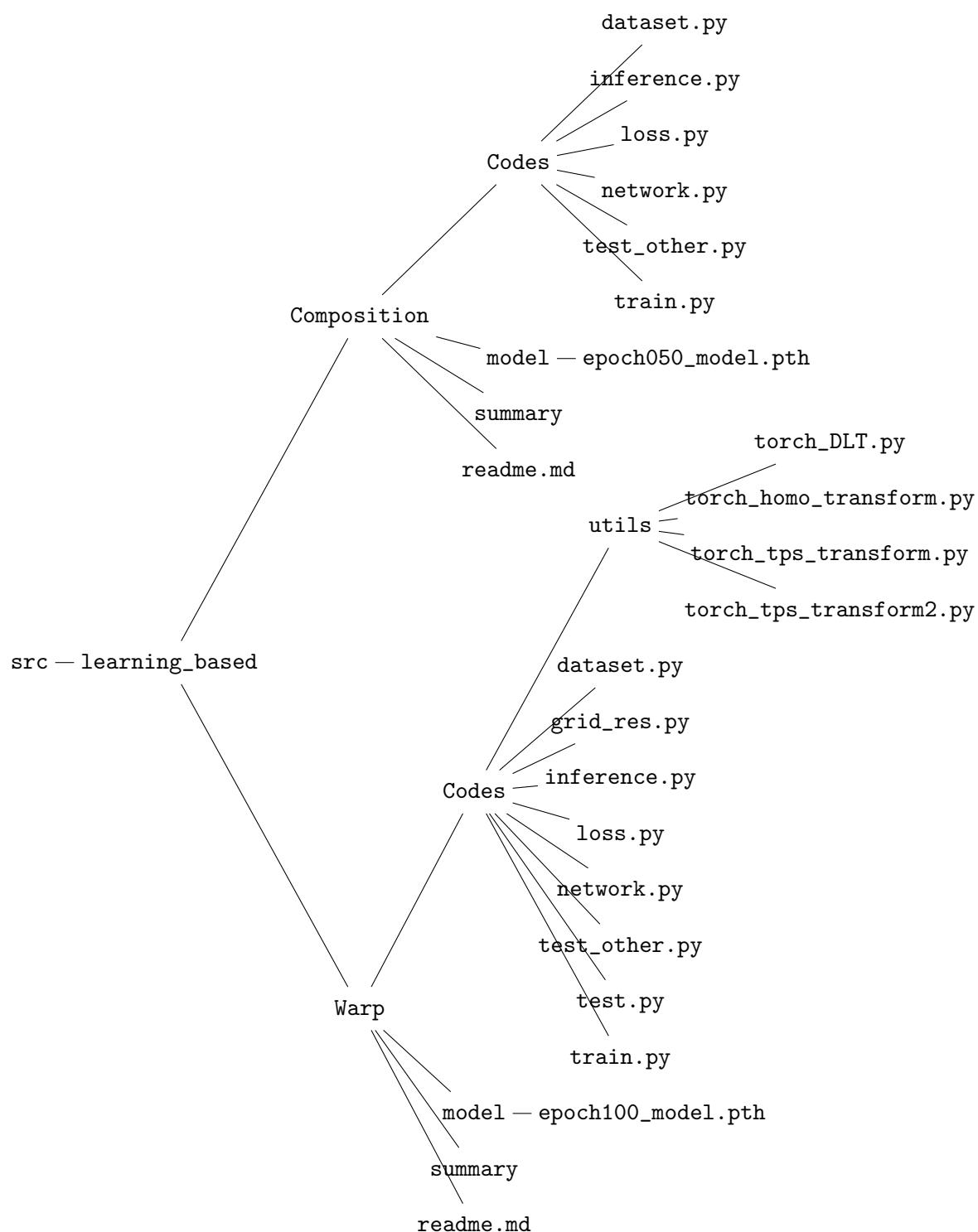
python Panorama.py hoặc python3 Panorama.py

3. Sau khi chương trình thực thi thành công, tại thư mục output, chương trình sẽ tạo ra hai ảnh:

- mapped_image: là ảnh khớp nối các đặc trưng từ kỹ thuật SIFT.
- panorama_image: là hình ảnh ghép hoàn chỉnh cuối cùng.

6.3 Phương pháp học sâu

Cấu trúc tổ chức mã nguồn phương pháp học sâu:



Chi tiết các bước huấn luyện và thử nghiệm:

1. Dièu hướng đến thư mục `src` và sau đó vào thư mục `learning_based`.
2. Quá trình thử nghiệm được chia thành hai giai đoạn chính, "Warp" và "Composition", mỗi giai đoạn được tổ chức trong các thư mục tương ứng với tên gọi của chúng.
3. **Giai đoạn "Warp":**

Truy cập vào thư mục "Warp".

Huấn luyện mô hình học sâu:

- Truy cập thư mục **Codes**.
- Điều chỉnh đường dẫn dữ liệu thử nghiệm trong tệp **train.py** sao cho phù hợp với cấu trúc thư mục chính xác.
- Chạy tệp **train.py** để tiến hành huấn luyện mô hình.

Thực nghiệm với mô hình đã được huấn luyện trước:

- Tải tệp trọng số đã được huấn luyện trước và lưu vào thư mục **model**. Đường dẫn tải tệp trọng số có sẵn trong tệp **model.txt** hoặc có thể tải trực tiếp tại: https://drive.google.com/file/d/1GBwB0y3tUUs0YHErSqxDxoC_0m3BJUEt/view.
- Truy cập vào thư mục **Codes**. Điều chỉnh đường dẫn dữ liệu thử nghiệm trong tệp **inference.py** để đảm bảo chính xác.
- Chạy tệp **inference.py** để thực hiện suy luận của mô hình. Sau khi hoàn thành, chương trình sẽ tạo ra:
 - Một thư mục mới tên **ave_fusion**, chứa hình ảnh đầu ra của giai đoạn "Warp".
 - Các thư mục **mask1**, **mask2**, **warp1**, và **warp2**, được đặt cùng cấp với thư mục dữ liệu thử nghiệm. Những thư mục này hỗ trợ cho giai đoạn tiếp theo - giai đoạn "Composition".

4. Giai đoạn "Composition":

Truy cập vào thư mục "Composition".

Huấn luyện mô hình học sâu:

- Truy cập thư mục **Codes**.
- Điều chỉnh đường dẫn dữ liệu thử nghiệm trong tệp **train.py** sao cho phù hợp với cấu trúc thư mục của bạn.
- Chạy tệp **train.py** để tiến hành huấn luyện mô hình.

Thực nghiệm với mô hình đã được huấn luyện trước:

- Tải tệp trọng số đã được huấn luyện trước và lưu vào thư mục **model**. Đường dẫn tải tệp trọng số có sẵn trong tệp **model.txt** hoặc có thể tải trực tiếp tại: https://drive.google.com/file/d/10aG0ayEwRPhKVV_0wQwwHDFHC26iv30/view.
- Truy cập vào thư mục **Codes**. Điều chỉnh đường dẫn dữ liệu thử nghiệm trong tệp **inference.py** để đảm bảo chính xác.
- Chạy tệp **inference.py** để thực hiện suy luận của mô hình. Sau khi hoàn thành, chương trình sẽ tạo ra:
 - Trong quá trình suy luận của mô hình, chương trình tạo ra các thư mục mới tên **learn_mask1**, **learn_mask2**.
 - Chương trình tạo ra một thư mục mới cùng cấp tên **composition** chứa hình ảnh đã ghép hoàn chỉnh. Đây là ảnh kết quả của quá trình.

7 Kết quả và đánh giá

7.1 Các thang đo

Nhóm sử dụng hai thang đo là PSNR và SSIM.

1. PSNR (Peak Signal-to-Noise Ratio)
 PSNR là một chỉ số dùng để đo lường sự khác biệt giữa hai ảnh, thường được sử dụng để đánh giá chất lượng của ảnh tái tạo so với ảnh gốc. PSNR càng cao, chất lượng ảnh càng tốt và sự khác biệt giữa ảnh gốc và ảnh tái tạo càng nhỏ.

Công thức tính PSNR:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right)$$

Trong đó:

- MAX_I là giá trị lớn nhất có thể của một pixel.
- MSE là **Mean Squared Error** (Sai số bình phương trung bình) giữa ảnh gốc và ảnh tái tạo, được tính theo công thức:

$$\text{MSE} = \frac{1}{N \cdot M} \sum_{i=1}^N \sum_{j=1}^M (I(i, j) - K(i, j))^2$$

Trong đó:

- $I(i, j)$ và $K(i, j)$ là giá trị pixel của ảnh gốc và ảnh tái tạo tại tọa độ (i, j) .
- N, M là chiều cao và chiều rộng của ảnh.

2. SSIM (Structural Similarity Index)

SSIM là một chỉ số đo lường sự tương đồng cấu trúc giữa hai ảnh. Dựa trên ba yếu tố: độ sáng, độ tương phản và cấu trúc của ảnh để đo lường mức độ tương tự. SSIM có giá trị từ -1 đến 1, trong đó 1 đại diện cho sự tương đồng hoàn toàn giữa ảnh gốc và ảnh tái tạo.

Công thức tính SSIM:

$$\text{SSIM}(x, y) = [l(x, y)^\alpha] \cdot [c(x, y)^\beta] \cdot [s(x, y)^\gamma]$$

Trong đó:

- $l(x, y)$: hàm so sánh độ sáng.
- $c(x, y)$: hàm so sánh độ tương phản.
- $s(x, y)$: hàm so sánh cấu trúc.
- α, β, γ : các trọng số (thường đặt là 1).
- Các hàm $l(x, y), c(x, y), s(x, y)$ được định nghĩa dựa trên giá trị trung bình, độ lệch chuẩn, và hiệp phương sai (covariance) của các pixel trong các cửa sổ nhỏ (thường là 8×8 hoặc 11×11) trượt trên hai ảnh.

7.2 Phương pháp truyền thống

7.2.1 Kết quả thử nghiệm

Nhóm tiến thành thử nghiệm trên tám trường hợp khác nhau từ tập dữ liệu thử nghiệm của **UDIS-D**.



Hình 41: Input 1



Hình 42: Input 2



Hình 43: Output



Hình 44: Input 1

Hình 45: Input 2

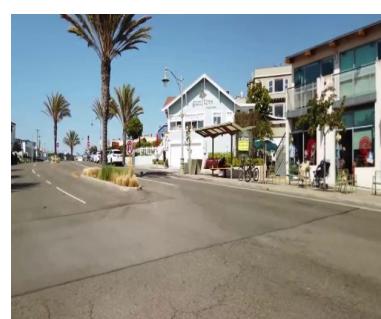


Hình 46: Output



Hình 47: Input 1

Hình 48: Input 2

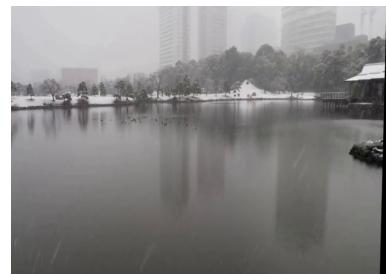


Hình 49: Output



Hình 50: Input 1

Hình 51: Input 2

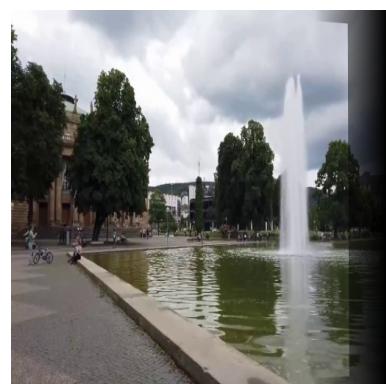


Hình 52: Output



Hình 53: Input 1

Hình 54: Input 2



Hình 55: Output

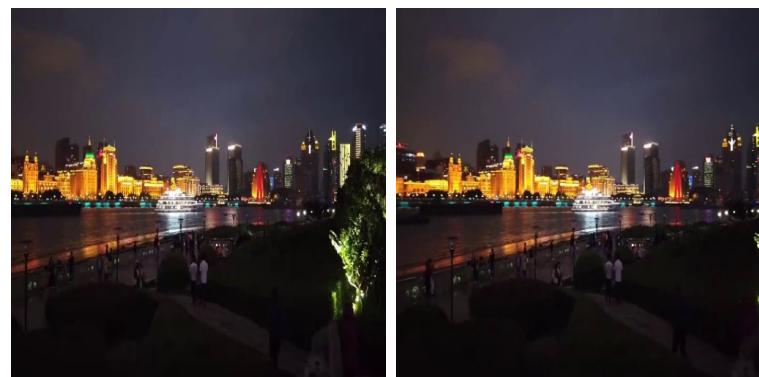


Hình 56: Input 1

Hình 57: Input 2

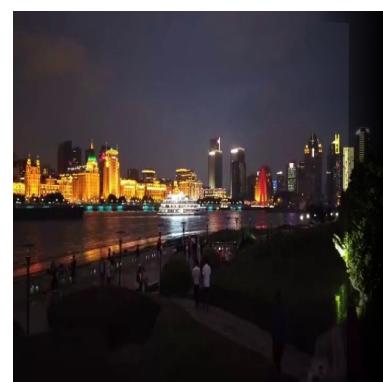


Hình 58: Output



Hình 59: Input 1

Hình 60: Input 2



Hình 61: Output



Hình 62: Input 1



Hình 63: Input 2



Hình 64: Output

7.2.2 Cài đặt thang đo

- **Mục tiêu:** Đánh giá chất lượng ghép ảnh bằng cách so sánh ảnh *input1* gốc và ảnh *input2* sau khi đã được "kéo giãn"(warp) sao cho khớp với *input1* trong vùng chồng lấp.

- **Input:**

- *Ảnh input1:* Ảnh gốc, chưa qua xử lý, đóng vai trò là ảnh tham chiếu.
- *Ảnh input2:* Ảnh gốc thứ hai.
- *Homography matrix H:* Một ma trận 3×3 , được tính toán từ các điểm đặc trưng (feature points) tương đồng giữa hai ảnh, chứa thông tin về cách "kéo giãn"(warp) *input2* sao cho khớp với *input1*.

- **Các bước thực hiện:**

1. Warp ảnh *input2*:

- Sử dụng *Homography matrix H* để biến đổi (warp) ảnh *input2*, tạo ra ảnh *warp2*, sao cho các điểm đặc trưng trên *input2* khớp với các điểm tương ứng trên *input1*.

2. Xác định vùng chồng lấp (*Overlap Region*):

- Warp ảnh *input2* ở dạng ảnh xám để xác định vùng chồng lấp.
- Pixel có giá trị lớn hơn 0 sau khi warp tạo thành vùng chồng lấp, được lưu dưới dạng ảnh nhị phân *mask*.

- Resize *mask* về kích thước của *input1* và nhân lên 3 kênh để phù hợp với ảnh màu (RGB).
3. Giữ lại phần *input1* trong vùng chồng lấp:
 - Nhân (*bitwise AND*) ảnh *input1* với *mask*.
 - Kết quả là ảnh *masked_img1*, trong đó chỉ còn lại phần hình ảnh của *input1* nằm trong vùng chồng lấp.
 4. Giữ lại phần *warp2* trong vùng chồng lấp:
 - Warp lại ảnh *input2* với kích thước của *input1* và nhân (*bitwise AND*) ảnh *warp2* với *mask*.
 - Kết quả là ảnh *masked_warp2*, trong đó chỉ còn lại phần hình ảnh của *warp2* trong vùng chồng lấp.
 5. Tính toán PSNR và SSIM:
 - Tính PSNR và SSIM giữa *masked_img1* (phần *input1* trong vùng chồng lấp) và *masked_warp2* (phần *input2* đã warp, trong vùng chồng lấp).
 - PSNR đo độ sai lệch về giá trị pixel giữa hai ảnh. PSNR càng cao, *masked_warp2* càng giống với *masked_img1*.
 - SSIM đo độ tương đồng về cấu trúc giữa hai ảnh. SSIM càng gần 1, *masked_warp2* càng giống với *masked_img1* về mặt cấu trúc.

• **Giải thích:**

- Bằng cách so sánh *masked_img1* và *masked_warp2*, ta đánh giá được mức độ khớp giữa *input2* sau khi warp và *input1* trong vùng chồng lấp.
- *mask* giúp tập trung vào vùng chồng lấp, bỏ qua phần ảnh không liên quan.
- Việc resize *mask* và chuyển về 3 kênh đảm bảo phép nhân (*bitwise AND*) đúng cho ảnh màu.

7.2.3 Kết quả thang đo

Nhóm tiến hành tính toán từng chỉ số PSNR và SSIM của tám trường hợp thử nghiệm phía trên và các chỉ số trung bình. Kết quả thang đo được thống kê như sau:

Test Case	PSNR	SSIM
1	6.16	0.51
2	6.49	0.55
3	5.92	0.48
4	5.68	0.48
5	6.58	0.56
6	5.80	0.47
7	6.62	0.57
8	6.00	0.51
Average	6.15625	0.51625

Bảng 1: PSNR và SSIM

Nhận xét chung:

- Giá trị PSNR trung bình là 6.15625, đây là một giá trị **khá thấp**. Điều này cho thấy ảnh ghép (masked_warp2) có sự sai lệch đáng kể về giá trị pixel so với ảnh gốc (masked_img1) trong vùng chồng lấp.
- Giá trị SSIM trung bình là 0.51625, cũng là một giá trị **tương đối thấp**. Điều này cho thấy ảnh ghép và ảnh gốc có sự khác biệt về cấu trúc, độ sáng và độ tương phản trong vùng chồng lấp.
- Có sự biến động trong các giá trị PSNR và SSIM giữa các trường hợp, với PSNR dao động từ 5.68 đến 6.62 và SSIM từ 0.47 đến 0.57.

Phân tích PSNR:

- Như đã đề cập, giá trị **PSNR trung bình thấp** (6.15625) cho thấy phương pháp feature-based gặp khó khăn trong việc tái tạo chính xác giá trị pixel của ảnh input2 sau khi warp trong vùng chồng lấp.
- Sự chênh lệch pixel này có thể do nhiều nguyên nhân:
 - Sai số trong việc tính toán homography:** Ma trận homography có thể không hoàn hảo, dẫn đến việc warp ảnh input2 không hoàn toàn khớp với input1.
 - Hiện tượng biến dạng (distortion):** Quá trình warp luôn đi kèm với một mức độ biến dạng nhất định, đặc biệt là khi hai ảnh có sự khác biệt lớn về phôi cảnh.
 - Khác biệt về điều kiện chụp ảnh:** Hai ảnh input1 và input2 có thể được chụp trong điều kiện ánh sáng, phoi sáng khác nhau, dẫn đến sự khác biệt về giá trị pixel ngay cả trong vùng chồng lấp.
 - Lỗi làm tròn của phép nhân mask:** Khi nhân với mask nhị phân, các giá trị pixel có thể bị thay đổi ít nhiều.
- Trường hợp 4 có **PSNR thấp nhất** (5.68), có thể do một trong các nguyên nhân trên gây ra sự sai lệch lớn về giá trị pixel.
- Trường hợp 5 và 7 có **PSNR cao nhất** (lần lượt là 6.58 và 6.62), cho thấy việc ghép ảnh hoạt động tốt hơn trong hai trường hợp này, có thể do ít biến dạng hơn hoặc homography chính xác hơn.

Phân tích SSIM:

- Giá trị **SSIM trung bình** 0.51625 cho thấy phương pháp feature-based chưa thực sự hiệu quả trong việc bảo toàn cấu trúc, độ sáng và độ tương phản của ảnh trong vùng chồng lấp.
- Điều này có thể do:
 - Hiện tượng mờ (blurring):** Quá trình warp và blending có thể làm mờ ảnh, làm giảm độ sắc nét và mất chi tiết, dẫn đến SSIM thấp.
 - Sự không nhất quán về độ sáng và độ tương phản:** Do input1 và input2 có thể được chụp trong điều kiện khác nhau, vùng chồng lấp sau khi ghép có thể bị thay đổi đột ngột về độ sáng và độ tương phản, làm giảm SSIM.

- **Biến dạng cấu trúc:** Việc warp ảnh có thể làm biến dạng các đường nét, cấu trúc trong ảnh, ảnh hưởng đến SSIM.
- Trường hợp 6 có **SSIM thấp nhất** (0.47), có thể do ảnh bị mờ hoặc biến dạng cấu trúc nghiêm trọng sau khi warp.
- Trường hợp 5 và 7 có **SSIM cao nhất** (lần lượt là 0.56 và 0.57), cho thấy hai trường hợp này bảo toàn cấu trúc ảnh tốt hơn sau khi ghép.

7.3 Phương pháp học sâu

7.3.1 Kết quả thử nghiệm

Nhóm tiến hành thử nghiệm trên tám trường hợp khác nhau từ tập dữ liệu thử nghiệm của **UDIS-D**.



Hình 65: Input 1

Hình 66: Input 2



Hình 67: Output



Hình 68: Input 1

Hình 69: Input 2



Hình 70: Output



Hình 71: Input 1

Hình 72: Input 2



Hình 73: Output



Hình 74: Input 1

Hình 75: Input 2

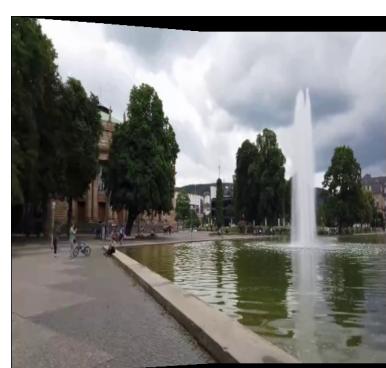


Hình 76: Output



Hình 77: Input 1

Hình 78: Input 2



Hình 79: Output



Hình 80: Input 1



Hình 81: Input 2



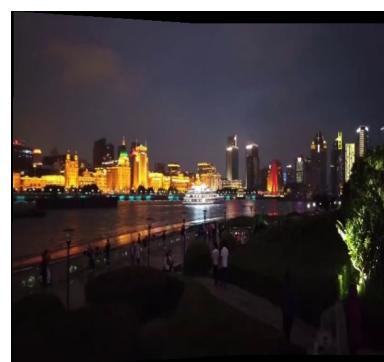
Hình 82: Output



Hình 83: Input 1



Hình 84: Input 2



Hình 85: Output



Hình 86: Input 1



Hình 87: Input 2



Hình 88: Output

7.3.2 Cài đặt thang đo

- **Mục tiêu:** Dánh giá chất lượng ghép ảnh bằng cách so sánh ảnh *input1* gốc và ảnh *input2* sau khi đã được “biến đổi” (warp) bằng mô hình deep learning sao cho khớp với *input1* trong vùng chồng lấp.

- **Input:**

- *Ảnh input1*: Ảnh gốc, chưa qua xử lý, đóng vai trò là ảnh tham chiếu.
- *Ảnh warp_mesh*: Ảnh *input2* đã được “biến đổi” (warp) bằng mô hình deep learning để khớp với *input1*.
- *warp_mesh_mask*: Mặt nạ nhị phân cho biết phần nào của ảnh *warp_mesh* là hợp lệ (phần mà ảnh *input2* được warp đến).

- **Các bước thực hiện:**

1. Giữ lại phần *input1* trong vùng chồng lấp:
 - Nhân ảnh *input1* với *warp_mesh_mask*.
 - Kết quả là ảnh *input1 · warp_mesh_mask*, trong đó chỉ còn lại phần hình ảnh của *input1* tương ứng với vùng mà *input2* được warp đến.
2. Giữ lại phần *warp_mesh* trong vùng chồng lấp:
 - Nhân ảnh *warp_mesh* với *warp_mesh_mask*.
 - Kết quả là ảnh *warp_mesh · warp_mesh_mask*, trong đó chỉ còn lại phần hình ảnh của *warp_mesh* tương ứng với vùng mà *input2* được warp đến.

3. Tính toán PSNR và SSIM:

- Tính PSNR và SSIM giữa hai ảnh $input1 \cdot warp_mesh_mask$ và $warp_mesh \cdot warp_mesh_mask$.
- PSNR đo độ sai lệch về giá trị pixel giữa hai ảnh. PSNR càng cao, $warp_mesh$ càng giống với $input1$ về mặt giá trị pixel trong vùng chồng lấp.
- SSIM đo độ tương đồng về cấu trúc (độ sáng, độ tương phản, các chi tiết) giữa hai ảnh. SSIM càng gần 1, $warp_mesh$ càng giống với $input1$ về mặt cấu trúc trong vùng chồng lấp.

- **Giải thích:**

- Bằng cách so sánh $input1 \cdot warp_mesh_mask$ và $warp_mesh \cdot warp_mesh_mask$, ta đánh giá được $input2$ sau khi được “biến đổi” bởi mô hình deep learning khớp với $input1$ tốt như thế nào trong vùng chồng lấp.
- $warp_mesh_mask$ đóng vai trò quan trọng, giúp chúng ta chỉ tập trung vào vùng ảnh mà $input2$ được warp đến, bỏ qua phần ảnh không liên quan.

7.3.3 Kết quả thang đo

Nhóm tiến hành tính toán từng chỉ số PSNR và SSIM của tám trường hợp thử nghiệm phía trên và các chỉ số trung bình. Kết quả thang đo được thống kê như sau:

Test Case	PSNR	SSIM
1	25.320972	0.936489
2	26.452366	0.937623
3	26.956150	0.952685
4	24.932880	0.890710
5	25.945891	0.858960
6	23.719968	0.868097
7	28.407929	0.936679
8	33.621817	0.966552
Average	26.92	0.919

Bảng 2: PSNR và SSIM

Phân tích PSNR:

- Giá trị PSNR cao nhất:** 33.62 tại Test Case 8. Đây là chỉ số PSNR cao nhất trong tất cả các test case, cho thấy chất lượng ảnh tại test case này tốt nhất trong tập dữ liệu.
- Giá trị PSNR thấp nhất:** 23.72 tại Test Case 6. Đây là test case có PSNR thấp nhất, điều này chỉ ra rằng sự khác biệt giữa ảnh gốc và ảnh tái tạo tại đây khá lớn, dẫn đến chất lượng ảnh thấp.
- Giá trị trung bình PSNR:** 26.92, cho thấy tổng thể chất lượng ảnh tái tạo trong tập dữ liệu là khá tốt, mặc dù có một số test case có PSNR thấp hơn.
- Các giá trị PSNR tương đối gần nhau, cho thấy rằng các ảnh tái tạo đều có chất lượng tương đương, với sự dao động nhỏ.

Phân tích SSIM:

- **Giá trị SSIM thấp nhất:** 0.858960 tại Test Case 5. Đây là test case có SSIM thấp nhất, chỉ ra rằng sự tương đồng cấu trúc giữa ảnh gốc và ảnh tái tạo tại đây khá thấp.
- **Giá trị trung bình SSIM:** 0.919, cho thấy tổng thể chất lượng cấu trúc của ảnh tái tạo trong tập dữ liệu là khá tốt. Tuy nhiên, vẫn có một số test case có SSIM thấp hơn, cho thấy có sự biến động trong sự tương đồng cấu trúc giữa các ảnh.
- Các giá trị SSIM có sự dao động giữa các test case, từ mức cao 0.966 đến mức thấp 0.858960. Điều này chỉ ra rằng một số ảnh tái tạo rất giống với ảnh gốc về cấu trúc, trong khi một số khác có sự khác biệt rõ rệt về cấu trúc. Các test case có SSIM cao thể hiện chất lượng tái tạo ảnh tốt, trong khi các test case có SSIM thấp cần cải thiện về mặt cấu trúc.

7.4 So sánh giữa hai phương pháp

7.4.1 Về giá trị PSNR

- **Phương pháp learning-based vượt trội hơn đáng kể so với feature-based.** Phương pháp learning-based đạt PSNR trung bình là 26.92, cao hơn rất nhiều so với 6.15625 của feature-based.
- **PSNR của phương pháp learning-based ổn định hơn,** dao động từ 23.719968 đến 33.621817, trong khi feature-based có khoảng dao động lớn hơn (5.68 đến 6.62).
- **Kết quả PSNR cao của learning-based** cho thấy ảnh *warp_mesh* (ảnh *input2* sau khi warp) có độ sai lệch pixel rất thấp so với ảnh *input1* trong vùng chồng lấp. Điều này chứng tỏ mô hình deep learning đã học được cách "kéo giãn" ảnh *input2* một cách chính xác để khớp với *input1*.

7.4.2 Về giá trị SSIM

- **Phương pháp learning-based cũng vượt trội hơn so với feature-based.** Learning-based đạt SSIM trung bình là 0.919, cao hơn đáng kể so với 0.51625 của feature-based.
- **SSIM của learning-based ổn định hơn,** dao động từ 0.858960 đến 0.966552, trong khi feature-based có khoảng dao động lớn hơn (0.47 đến 0.57).
- **Kết quả SSIM cao của learning-based** cho thấy ảnh *warp_mesh* không chỉ khớp với *input1* về giá trị pixel mà còn duy trì được sự tương đồng về cấu trúc, độ sáng và độ tương phản. Điều này chứng tỏ mô hình deep learning đã hiểu được nội dung của ảnh và thực hiện warp một cách thông minh, giữ lại các chi tiết quan trọng.

7.4.3 Phân tích dựa trên ảnh kết quả

Feature-based:

- Ảnh kết quả có hiện tượng mờ (*blurring*) và biến dạng (*distortion*), đặc biệt ở vùng rìa và các khu vực nhiều chi tiết.

- Một số ảnh xuất hiện "bóng mờ" (*ghosting*) do sự không nhất quán trong ghép ảnh.
- Đường thẳng bị bẻ cong, chi tiết bị mờ, và khác biệt rõ rệt về độ sáng và màu sắc trong vùng chồng lấp.

Learning-based:

- Ảnh kết quả liền mạch hơn, hầu như không có hiện tượng "ghosting".
- Các chi tiết được giữ lại tốt, và sự chuyển tiếp giữa hai ảnh trong vùng chồng lấp rất mượt mà.
- Độ sáng và màu sắc được bảo toàn tốt hơn. Tuy nhiên, một số ảnh vẫn còn hiện tượng hơi mờ (*blur*) nhẹ, cho thấy khả năng cải thiện của mô hình.

7.4.4 Ưu nhược điểm của từng phương pháp

Feature-based:

- **Ưu điểm:** Dễ hiểu, dễ cài đặt, không yêu cầu dữ liệu huấn luyện.
- **Nhược điểm:** Hiệu suất kém hơn, dễ bị ảnh hưởng bởi biến dạng và khác biệt trong điều kiện chụp, khó xử lý thay đổi lớn về phối cảnh, kết quả phụ thuộc vào chất lượng trích xuất đặc trưng và tính toán homography.

Learning-based:

- **Ưu điểm:** Hiệu suất cao hơn, xử lý tốt các trường hợp phức tạp, kết quả ghép ảnh tự nhiên và liền mạch.
- **Nhược điểm:** Phức tạp hơn, đòi hỏi dữ liệu huấn luyện lớn, thời gian huấn luyện lâu, yêu cầu tài nguyên tính toán lớn.

7.4.5 Kết luận

Dựa trên kết quả định lượng (PSNR, SSIM) và đánh giá trực quan, phương pháp learning-based vượt trội hơn hẳn so với feature-based. Mô hình deep learning đã học được cách warp ảnh một cách thông minh, giữ lại chi tiết quan trọng và tạo ra ảnh ghép tự nhiên. Tuy nhiên, learning-based có nhược điểm về độ phức tạp và yêu cầu dữ liệu huấn luyện.

8 Kết luận và hướng phát triển

8.1 Kết luận

Bài khảo sát đã cung cấp một cái nhìn toàn diện về lĩnh vực ghép ảnh (Image Stitching), từ các phương pháp truyền thống đến những tiến bộ gần đây dựa trên học sâu. Các nghiên cứu cho thấy, sự phát triển của Trí tuệ nhân tạo, đặc biệt là các mô hình học sâu, không

chỉ cải tiến hiệu suất xử lý mà còn tạo ra một cuộc cách mạng trong Thị giác máy tính, góp phần nâng cao độ chính xác và chất lượng của quy trình ghép ảnh.

Những mô hình học sâu hiện đại đã giải quyết hiệu quả nhiều thách thức tồn tại trong các phương pháp truyền thống. Đặc biệt, chúng cải thiện đáng kể các vấn đề như hiện tượng bóng mờ và hình ảnh kép trong vùng chồng lấn giữa các ảnh, cũng như giảm thiểu sự khác biệt về độ sâu và thị sai trong các cảnh ba chiều phức tạp. Bên cạnh đó, các mô hình này cũng cho thấy tiềm năng trong việc xử lý biến dạng hình học và giảm thiểu sự khác biệt về ánh sáng và màu sắc giữa các hình ảnh đầu vào.

Ghép ảnh không chỉ là một bài toán kỹ thuật quan trọng mà còn là nền tảng cho nhiều ứng dụng thực tiễn. Những tiến bộ trong lĩnh vực này không chỉ góp phần cải thiện chất lượng hình ảnh mà còn mở ra cơ hội phát triển cho nhiều lĩnh vực khác, như thực tế ảo, y tế, an ninh, và bản đồ số. Đây vẫn là một hướng nghiên cứu tiềm năng với nhiều triển vọng trong tương lai.

8.2 Hạn chế

Mặc dù đã đạt được nhiều tiến bộ, lĩnh vực ghép ảnh vẫn phải đối mặt với những thách thức đáng kể. Một trong những vấn đề lớn là khả năng xử lý các cảnh phức tạp chứa độ nhiễu cao, chẳng hạn như ảnh bị mờ, mất chi tiết, hoặc ảnh hưởng bởi điều kiện ánh sáng kém. Những tình huống này thường làm giảm chất lượng ghép nối và đòi hỏi các thuật toán phải mạnh mẽ hơn để đảm bảo tính chính xác và đồng nhất.

Bên cạnh đó, sự di chuyển không đồng bộ của các đối tượng trong cảnh vẫn là một thách thức chưa thể giải quyết triệt để. Điều này đặc biệt phổ biến trong các bối cảnh năng động, như đường phố đông đúc, nơi mà các xe cộ hoặc người đi bộ thay đổi vị trí liên tục giữa các khung hình. Sự không nhất quán này gây khó khăn trong việc căn chỉnh và dán đến hiện tượng bóng mờ hoặc hình ảnh kép.

Ngoài ra, sự khác biệt về độ sâu và thị sai giữa các đối tượng trong cảnh ba chiều phức tạp vẫn chưa được xử lý hoàn toàn. Các đối tượng ở khoảng cách khác nhau thường xuất hiện không đồng nhất trong các ảnh đầu vào, đòi hỏi các thuật toán phải cải tiến hơn nữa để đảm bảo sự chính xác trong việc căn chỉnh và tái tạo không gian ba chiều.

8.3 Hướng phát triển tương lai

Để tiếp tục phát triển, việc kết hợp các kỹ thuật học sâu với tri thức miền (domain knowledge) và các giải pháp tối ưu hóa. Điều này không chỉ giúp xử lý dữ liệu đa dạng hơn mà còn tăng cường khả năng áp dụng trong các bài toán phức tạp như y tế, an ninh và bản đồ số. Sự kết hợp này hứa hẹn sẽ cải thiện độ chính xác và hiệu quả, đặc biệt khi phải làm việc với các tập dữ liệu lớn và đa dạng.

Một hướng nghiên cứu khác là giải quyết triệt để các vấn đề liên quan đến biến dạng hình học và thị sai. Đây là những thách thức phổ biến trong các cảnh ba chiều phức tạp, đặc biệt khi các hình ảnh được chụp từ các góc độ khác nhau hoặc các khoảng cách không đồng nhất.

Ngoài ra, việc kết hợp kỹ thuật siêu phân giải với Image Stitching cũng đang là một xu hướng đáng chú ý. Phương pháp này không chỉ nâng cao độ phân giải của hình ảnh toàn cảnh mà còn cải thiện chất lượng và chi tiết, đáp ứng nhu cầu ngày càng cao trong các ứng

dụng như hình ảnh vệ tinh, y tế, và thực tế ảo.

Các nghiên cứu hiện nay cũng tập trung vào phát triển các mô hình học sâu không giám sát, nhằm giảm sự phụ thuộc vào dữ liệu được gán nhãn thủ công. Điều này không chỉ giúp tăng cường khả năng tổng quát hóa của các mô hình mà còn mở rộng phạm vi áp dụng trong các ngữ cảnh mới và ít dữ liệu. Đồng thời, việc tích hợp các phương pháp dựa trên đặc trưng truyền thống với học sâu cũng được chú ý, nhằm tận dụng ưu điểm của cả hai cách tiếp cận để đạt hiệu suất cao hơn.

Cuối cùng, khả năng xử lý thời gian thực là một yếu tố quan trọng cần được cải thiện trong tương lai. Các thuật toán và mô hình mới đang hướng tới việc tăng tốc độ xử lý để đáp ứng nhu cầu của các ứng dụng đòi hỏi tính tức thời, chẳng hạn như giám sát an ninh, thực tế ảo, và các hệ thống tương tác trực tiếp.

9 Tài liệu tham khảo

Tài liệu

- [1] E. Adel, M. Elmogy, and H. Elbakry. Image Stitching based on Feature Extraction Techniques: A Survey. *International Journal of Computer Applications*, vol. 99, no. 6, pp. 0975 – 8887 (2014).
- [2] L. Wei, Z. Zhong, C. Lang, and Z. Yi. A survey on image and video stitching. *Virtual Reality & Intelligent Hardware*, vol. 1, no. 1, pp. 55 – 83 (2019).
- [3] M. Fu, H. Liang, C. Zhu, Z. Dong, R. Sun, and Y. Yue. Image Stitching Techniques Applied to Plane or 3-D Models: A Review. *IEEE Sensors Journal*, vol. 23, no. 8, pp. 8060 - 8079 (2023).
- [4] D. Ghosh and N. Kaabouch. A survey on image mosaicing techniques. *Journal of Visual Communication and Image Representation*, vol. 34, pp. 1 – 11 (2016).
- [5] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao. Parallax-Tolerant Unsupervised Deep Image Stitching. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7399–7408 (2023).
- [6] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao. Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images. *IEEE Transactions on Image Processing*, vol. 30, pp. 6184–6197 (2021).
- [7] D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. Camera-ready version for CVPR 2018 Deep Learning for Visual SLAM Workshop (DL4VSLAM2018).
- [8] N. Yan, Y. Mei, L. Xu, H. Yu, B. Sun, Z. Wang, and Y. Chen. Deep Learning on Image Stitching With Multi-viewpoint Images: A Survey. *Neural Processing Letters*, vol. 55, pp. 3863–3898, Published: 23 March 2023.
- [9] E. Karami, S. Prasad, and M. Shehata. Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images. In *Proceedings of the 2015 Newfoundland Electrical and Computer Engineering Conference*, St. John's, Canada, November 2015.
- [10] G. Hong and Y. Zhang. Combination of feature-based and area-based image registration technique for high resolution remote sensing image. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, pp. 377-380 (2007), DOI: 10.1109/IGARSS.2007.4422809.
- [11] X. Tong, Z. Ye, Y. Xu, S. Gao, H. Xie, Q. Du, S. Liu, X. Xu, S. Liu, K. Luan, and U. Still. Image Registration With Fourier-Based Image Correlation: A Comprehensive Review of Developments and Applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 10, pp. 4062-4081 (2019), DOI: 10.1109/JSTARS.2019.2937690.
- [12] C. Qin and X. Ran. Efficient Unsupervised Image Stitching Using Attention Mechanism with Deep Homography Estimation. *Computers, Materials & Continua*, vol. 79, no. 1, pp. 1319-1334 (2024), DOI: 10.32604/cmc.2024.048850.

- [13] M. Kim, J. Lee, B. Lee, S. Im, and K. H. Jin. Implicit Neural Image Stitching With Enhanced and Blended Feature Reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4087-4096 (2024).