



ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC BÁCH KHOA  
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH



BÁO CÁO ĐỒ ÁN CHUYÊN NGÀNH

CẢI THIỆN ĐỊNH VỊ TRỰC QUAN BẰNG  
HƯỚNG TIẾP CẬN HỌC SÂU

HỘI ĐỒNG: Khoa học máy tính

GVHD: Nguyễn Đức Dũng

GVPB: GV phản biện

—o0o—

Sinh viên thực hiện:

Lê Minh Nghĩa MSSV: 2010445

Phạm Khai Anh Duy MSSV: 2011015

Nguyễn Trọng Nhân MSSV: 2011744

TP. HỒ CHÍ MINH, 12/2023

### **Lời cam đoan**

Chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng chúng tôi dưới sự hướng dẫn của TS. Nguyễn Đức Dũng. Nội dung nghiên cứu và các kết quả đều là trung thực và chưa từng được công bố trước đây. Các số liệu được sử dụng cho quá trình phân tích, nhận xét được chính chúng tôi thu thập từ nhiều nguồn khác nhau và sẽ được ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, chúng tôi cũng có sử dụng một số nhận xét, đánh giá và số liệu của các tác giả khác, cơ quan tổ chức khác. Tất cả đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kì sự gian lận nào, chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án tốt nghiệp của mình. Trường Đại học Bách Khoa Thành phố Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện.

### **Lời ngỏ**

Đồ án chuyên ngành được hoàn thành dưới sự hướng dẫn khoa học của TS. Nguyễn Đức Dũng, Khoa Khoa học và Kỹ Thuật Máy Tính, Trường Đại học Bách Khoa - Đại học Quốc gia Thành phố Hồ Chí Minh. Nhóm thực hiện xin chân thành cảm ơn TS. Nguyễn Đức Dũng, đã giúp đỡ chúng em kiến thức chuyên môn, thảo luận, đưa ra gợi ý và tạo điều kiện thuận lợi cho chúng em hoàn thành đồ án.

Chúng em cũng xin gửi lời cảm ơn đặc biệt đến quý thầy cô phản biện, những người đã đọc và đóng góp ý kiến để chúng em hoàn thiện đồ án chuyên ngành của mình. Nhóm thực hiện cũng xin bày tỏ lòng biết ơn sâu sắc đến quý thầy cô Khoa Khoa học và Kỹ Thuật Máy tính, Trường Đại học Bách Khoa - Đại học Quốc gia Thành phố Hồ Chí Minh, là những người đã truyền thụ kiến thức chuyên môn, đã tạo điều kiện cho chúng em học tập và phát triển trong suốt quãng thời gian vừa qua.

Kính chúc quý thầy cô sức khoẻ, thành công và tiếp tục đào tạo những thế hệ sinh viên mới trong tương lai.

Chúng em xin chân thành cảm ơn.

**Nhóm thực hiện**

Lê Minh Nghĩa  
Phạm Khai Anh Duy  
Nguyễn Trọng Nhân

## Tóm tắt nội dung

Khả năng định vị toàn cầu có vai trò cốt lõi trong những lĩnh vực phụ thuộc vào việc nhận biết và tương tác với môi trường xung quanh như xe tự hành, robot và công nghệ thực tế ảo AR. Trước đây, những công nghệ này sẽ phụ thuộc vào những hệ thống định vị toàn cầu như GPS. Tuy nhiên, những hệ thống này vẫn còn những giới hạn nhất định. Vì vậy nên, bài toán định vị trực quan - Visual Localization - đã được đưa ra nhằm đạt được kết quả chất lượng hơn, thông qua dữ liệu trực quan thu được tại vị trí đó.

Hai hướng tiếp cận đã có những cải tiến liên tục trong những năm gần đây chính là hướng nhận dạng địa điểm trực quan - Visual Place Recognition - và hướng ước tính vị trí của máy ảnh - Pose Estimation. Với hướng tiếp cận nhận dạng địa điểm trực quan, mô hình sẽ nhận vào một ảnh truy vấn và chọn ra một hay nhiều ảnh từ tập ảnh được cung cấp, thể hiện cùng một cảnh với ảnh đầu vào. Với hướng tiếp cận ước tính vị trí của máy ảnh, từ cặp ảnh truy vấn và ảnh tham khảo đầu vào, mô hình sẽ tính toán vị trí và hướng quay của máy ảnh trong không gian.

Nhận thấy rằng hai hướng tiếp cận của bài toán khi đứng riêng đều đã có kết quả khả quan trong phạm vi của mình, và dữ liệu đầu ra và đầu vào của hai hướng tiếp cận tương thích với nhau, nhóm quyết định sẽ kết hợp hai bài toán này thành một quy trình hoàn chỉnh, nhằm bổ trợ cho những khiếm khuyết của mỗi hướng. Cụ thể hơn:

- Với hướng tiếp cận nhận dạng địa điểm trực quan, nhóm sẽ sử dụng mô hình MixVPR. Đây sẽ là nửa đầu của quy trình, cung cấp khả năng mở rộng lên những không gian rộng lớn như thành phố.
- Với hướng tiếp cận ước tính vị trí máy ảnh, nhóm sẽ sử dụng hướng tiếp cận 2D-2D được đề xuất trong Map-free Relocalization. Đây sẽ là nửa còn lại của quy trình, cung cấp khả năng đưa ra một dự đoán cụ thể cho quy trình.

# Mục lục

<b>1 GIỚI THIỆU</b>	<b>1</b>
1.1 Động cơ nghiên cứu . . . . .	1
1.1.1 Nhu cầu sử dụng trong thực tế . . . . .	1
1.1.2 Những hướng đi đã được đề xuất . . . . .	2
1.2 Mục tiêu đề tài . . . . .	2
1.2.1 Khảo sát và phân tích những giải pháp đã có . . . . .	2
1.2.2 Tiến hành kiểm thử kết quả của phương pháp cơ sở . . . . .	3
1.2.3 Thiết kế giải pháp cải thiện hiệu quả của mô hình kết hợp . . . . .	3
1.2.4 Đánh giá và kiểm thử mô hình . . . . .	3
1.3 Phạm vi đề tài . . . . .	3
1.4 Cấu trúc đồ án chuyên ngành . . . . .	4
1.5 Kết chương . . . . .	4
<b>2 CÁC CÔNG TRÌNH LIÊN QUAN</b>	<b>6</b>
2.1 Những phương pháp đã được sử dụng . . . . .	6
2.1.1 Những phương pháp sử dụng biểu diễn 3D . . . . .	6
2.1.2 Những phương pháp hồi quy vị trí . . . . .	8
2.1.3 Những phương pháp khác . . . . .	9
2.1.4 Xác định những vấn đề hiện hữu trong bài toán định vị trực quan . . . . .	9
2.2 Nhận dạng địa điểm trực quan - Visual Place Recognition . . . . .	10
2.2.1 Học biểu diễn - Representational learning . . . . .	10
2.2.2 Học biểu diễn NetVLAD . . . . .	10
2.2.3 Tối ưu hóa đặc trưng - MAC . . . . .	10
2.2.4 Trung bình Hölder - Trung bình GeM . . . . .	10
2.2.5 Truy xuất và tái xếp hạng . . . . .	10
2.2.6 Học biểu diễn bằng Vision Transformer . . . . .	10
2.2.7 Học biểu diễn bằng Feature Mixer . . . . .	10
2.3 Ước tính vị trí của máy ảnh - Pose Estimation . . . . .	10
2.3.1 Hồi quy vị trí tuyệt đối - Absolute Pose Regression . . . . .	10
2.3.2 Hồi quy vị trí tương đối - Relative Pose Regression . . . . .	20
2.3.3 Tái tạo kiến trúc từ chuyển động - Structure From Motion . . . . .	24
2.4 Phân tích và tổng hợp . . . . .	24
2.5 Một số tập dữ liệu phổ biến được sử dụng . . . . .	24
2.5.1 Tập dữ liệu trong không gian nhỏ . . . . .	24
2.5.2 Tập dữ liệu thành thị . . . . .	26

<b>3 PHƯƠNG PHÁP ĐỀ XUẤT</b>	<b>28</b>
3.1 Tổng quan về mô hình được đề xuất . . . . .	28
3.1.1 MixVPR . . . . .	28
3.1.2 Mô hình tương quan 2D-2D của Map-free Relocalization . . . . .	28
3.2 Tiêu chí đánh giá . . . . .	28
3.2.1 MixVPR . . . . .	28
3.2.2 Mô hình tương quan 2D-2D của Map-free Relocalization . . . . .	28
<b>4 ĐO ĐẠC VÀ ĐÁNH GIÁ</b>	<b>29</b>
4.1 Mô hình MixVPR . . . . .	29
4.2 Mô hình Map-free Relocalization . . . . .	29
4.3 Hướng phát triển . . . . .	29
4.4 Kết luận . . . . .	29
<b>5 KẾ HOẠCH TƯƠNG LAI</b>	<b>30</b>
5.1 Thành quả đạt được . . . . .	30
5.2 Kế hoạch luận văn tốt nghiệp . . . . .	30

# Danh sách hình vẽ

2.1	Tổng quát về những phương pháp định vị trực quan quan trọng [16] . . . . .	6
2.2	Kịch bản khi mà tập dữ liệu huấn luyện bị hạn chế. Quỹ đạo của tập huấn luyện và tập kiểm thử có màu <b>đỏ</b> và <b>xanh lá</b> . Kết quả của những mô hình lấy vị trí của ảnh làm mốc là PoseNet [21] và MapNet [9] có quỹ đạo kết quả màu <b>xanh dương</b> và <b>tím</b> . Kết quả của phương pháp sử dụng biểu diễn 3D là Active Search [34] có quỹ đạo màu <b>xanh lam</b> . . . . .	8
2.3	Kiến trúc mô hình hồi quy vị trí tuyệt đối đơn ảnh [21] . . . . .	11
2.4	Minh họa mô hình CNN được áp dụng phân phôi Bernoulli [19] . . . . .	12
2.5	Minh họa kiến trúc mô hình LSTM PoseNet [45] . . . . .	12
2.6	Minh họa kiến trúc mô hình Hourglass PoseNet [25] . . . . .	13
2.7	Minh họa kiến trúc mô hình AtLoc [46] . . . . .	13
2.8	Minh họa kiến trúc mô hình APANet [13] . . . . .	14
2.9	Minh họa kiến trúc mô hình GPoseNet [20] . . . . .	14
2.10	Minh họa kiến trúc mô hình Multi-Scene Transformer [38, 39] . . . . .	15
2.11	Minh họa kiến trúc mô hình MapNet [9] . . . . .	16
2.12	Minh họa kiến trúc mô hình LSG [49] . . . . .	16
2.13	Minh họa kiến trúc mô hình VlocNet [44] . . . . .	17
2.14	Minh họa kiến trúc mô hình VlocNet++ [30] . . . . .	17
2.15	Minh họa kiến trúc mô hình DGRNet [23] . . . . .	18
2.16	Minh họa kiến trúc mô hình VidLoc [14] . . . . .	19
2.17	Minh họa kiến trúc mô hình NNet [22] . . . . .	21
2.18	Minh họa kiến trúc mô hình RelocNet [5] . . . . .	21
2.19	Minh họa kiến trúc mô hình CamNet [15] . . . . .	22
2.20	Minh họa kiến trúc mô hình EssNet [50] . . . . .	22
2.21	Minh họa kiến trúc mô hình Relative Neural Network [26] . . . . .	23
2.22	Minh họa kiến trúc mô hình AnchorNet [31] . . . . .	23
2.23	Minh họa kiến trúc mô hình hồi quy vị trí tương đối của Map-free [4] . . . . .	23
2.24	Minh họa tập dữ liệu 7-Scenes [40] . . . . .	25
2.25	Minh họa tập dữ liệu Cambridge Landmarks [21] . . . . .	25
2.26	Minh họa tập dữ liệu Niantic Map-free Relocalization [4] . . . . .	26
2.27	Minh họa tập dữ liệu Aachen Day-Night [37] . . . . .	26
2.28	Minh họa tập dữ liệu GSV-Cities [1] . . . . .	27
2.29	Minh họa tập dữ liệu San Francisco Extra Large [7] . . . . .	27

**Bảng những từ ngữ chuyên ngành được sử dụng và phiên bản tiếng Anh**

<b>STT</b>	<b>Bản tiếng Việt</b>	<b>Bản tiếng Anh</b>
1	Định vị trực quan	Visual Localization
2	Ước tính vị trí máy ảnh	Camera Pose Estimation
3	Hồi quy vị trí tuyệt đối	Absolute Pose Regression
4	Hồi quy vị trí tương đối	Relative Pose Regression
5	Máy ảnh lỗ kim	Pinhole Camera
6	Ma trận thiết yếu	Essential Matrix
7	Tìm sự tương ứng giữa đặc trưng ảnh	Feature Matching
8	Giải thuật 5 điểm ảnh	5-Point Solver
9	Thuật toán tính độ sâu ảnh qua một ảnh	Monocular Depth Estimation
10	Truy xuất ảnh	Image Retrieval
11	Bản đồ đám mây điểm 3D	3D point cloud
12	Tái tạo kiến trúc từ chuyển động	Structure from Motion
13	Mạng Nơ-ron tích chập	Convulated Neural Network
14	Trích xuất đặc trưng ảnh	Feature Extraction
15	Đặc trưng ảnh cục bộ	Local Descriptor
16	Đặc trưng ảnh toàn cục	Global Descriptor
17	Bản đồ đặc trưng ảnh	Feature Map
18	Lớp pha trộn đặc trưng ảnh	Feature Mixer
19	Lớp kết nối đầy đủ	Fully Connected Layer
20	Bộ nhớ dài-ngắn hạn	Long Short Term Memory
21	Hồi quy quá trình Gaussian suy luận biến phân ngẫu nhiên	Stochastic Variational Inference Gaussian Process Regressions - SVI GPs
22	Cơ chế tự tập trung	Self-Attention
23	Hàm kích hoạt	Activation Function
24	Chuẩn hóa trên mỗi lớp	Layer Normalization
25	Học có hỗ trợ	Auxiliary Learning
26	Đo lường cảm biến trực quan	Visual Odometry
27	Cục bộ hỗ trợ toàn cục	Local Support Global
28	Tổng hợp trung bình	Average Pooling
29	Entropy chéo	Cross-entropy
30	Lớp kết hợp kết nối đầy đủ	Fully-connected Fusion layer
31	Đơn vị đo quán tính	Inertial Measurement Unit
32	Hệ thống vệ tinh định vị toàn cầu	Global Navigation Satellite System
33	Cải thiện đồ thị vị trí	Pose graph optimization
34	Đồng thuận lân cận	Neighborhood Consensus
35	Hồi quy điểm trong cảnh	Scene Point Regression
36	Nhận diện địa điểm trực quan	Visual Place Recognition
37	Độ không đảm bảo phương sai đồng nhất	Homoscedastic uncertainty

**Bảng những từ viết tắt**

<b>STT</b>	<b>Bản viết tắt</b>	<b>Bản đầy đủ</b>
1	MLP	Multi-layer Perceptron
2	APR	Absolute Pose Regression
3	RPR	Relative Pose Regression
4	VPR	Visual Place Recognition
5	LSG	Local Support Global
6	FCFL	Fully-connected Fusion Layer
7	MSE	Mean square error
8	CTC	Cross transformation constraint
9	IMU	Inertial measurement unit
10	GNSS	Global Navigation Satellite System
11	VO	Visual Odometry
12	PGO	Pose graph optimization
13	NC	Neighborhood Consensus
14	SOTA	State-of-the-art
15	SfM	Structure-from-Motion
16	PnP	Perspective-n-Point
17	RANSAC	Random sample consensus

# Chương 1

## GIỚI THIỆU

*Nội dung chương 1 sẽ đề cập đến nội dung của bài toán định vị trực quan - Visual Localization, những giải pháp đã được đề xuất hiện nay của bài toán, và những hạn chế của chúng. Từ đó, nhóm sẽ xác định mục tiêu cần thực hiện và phạm vi của đề tài*

### 1.1 Động cơ nghiên cứu

#### 1.1.1 Nhu cầu sử dụng trong thực tế

Việc có thể nhận biết được môi trường xung quanh để có thể tương tác là một tác vụ cốt lõi trong những công nghệ được tích hợp vào cuộc sống hàng ngày của con người. Những công nghệ này bao gồm xe tự hành [12], robot [41], công nghệ tương tác thực tế ảo [27], định hướng [33], ...

Trước đây, những hệ thống định vị toàn cầu như GPS đã được sử dụng để xác định thông tin về môi trường. Tuy nhiên, những hệ thống này có một số khiếm khuyết như độ chính xác chỉ nằm trong khoảng vài mét, hiệu quả bị giới hạn ở không gian bên trong và thiếu thông tin về hướng quay nếu không sử dụng thêm la bàn. Để có thể đáp ứng nhu cầu về độ chính xác, bài toán định vị trực quan - Visual Localization - trong lĩnh vực thị giác máy tính đã được ra đời.

Bộ não con người có thể thực hiện bài toán định vị trực quan bằng trực giác. Tuy nhiên, để mô phỏng lại quá trình này, những giải thuật phức tạp liên quan đến việc xây dựng một cách biểu diễn phù hợp cho không gian như 3D-point cloud hay thực hiện feature matching đã được sử dụng. Những tác vụ này sẽ tiêu tốn rất nhiều tài nguyên để xây dựng và thực hiện.

Trong những năm gần đây, lĩnh vực này đã có những bước phát triển đáng kể, được thể hiện qua một lượng lớn bài báo nghiên cứu khoa học. Những bài báo này đã đưa ra những hướng đi đa dạng để giải quyết bài toán định vị trực quan, nhưng đa số đều tập trung vào hai chủ đề chính là:

- Tìm kiếm một cách biểu diễn tuy đơn giản, nhưng vẫn đảm bảo được tính hiệu quả trong việc truy xuất thông tin nhằm tiết kiệm tài nguyên để xây dựng, duy trì, mở rộng và sử dụng.
- Cải thiện độ chính xác của vị trí được truy xuất mà vẫn đảm bảo được tính hiệu quả.

## 1.1.2 Những hướng đi đã được đề xuất

Với mục tiêu là xác định được vị trí mà ảnh được chụp, nhiều phương pháp khác nhau đã được đề xuất để giải quyết bài toán định vị trực quan. Một nhóm phương pháp truyền thống mà cho đến hiện tại vẫn cho ra kết quả cạnh tranh là phương pháp dựa trên cấu trúc của cảnh - Structure-based Method. Phương pháp này sẽ dựa trên việc tái tạo lại cấu trúc của môi trường đang xét bằng một tập các điểm trong không gian 3D, tạo thành một 3D-point cloud để biểu diễn khu vực đang xét. Từ đó, tọa độ chính xác của vị trí chụp ảnh có thể được xác định. Để tối ưu hóa quá trình tìm kiếm tương quan 2D-3D, phương pháp truy xuất ảnh có thể được sử dụng để giới hạn lại không gian tìm kiếm [32]. Ngoài ra, ở bước xác định tương quan 2D-3D, thay vì sử dụng những phương pháp được định nghĩa sẵn bởi con người, mạng học sâu có thể được ứng dụng để trực tiếp xác định vị trí của các điểm ảnh trong không gian 3D [8].

Một hướng đi khác, sử dụng những ảnh có nét tương đồng với ảnh đầu vào, vị trí cuối cùng có thể được nội suy từ nhãn của những ảnh đó. Với phương pháp hồi quy tương đối, từ cặp ảnh gồm ảnh truy vấn và ảnh tham chiếu, mô hình sẽ xác định được khoảng cách về vị trí giữa hai ảnh [50]. Ngoài ra, thay vì truy xuất ảnh làm điểm mốc để xác định vị trí, phương pháp hồi quy vị trí tuyệt đối sẽ xây dựng cách biểu diễn của môi trường bên trong mô hình và có thể tính trực tiếp kết quả chỉ với đầu vào là ảnh truy vấn [21].

Trong đa số những phương pháp trên, tác vụ truy xuất ảnh đóng một vai trò quan trọng để có thể đạt được kết quả chính xác. Mục tiêu của quá trình truy xuất ảnh là để xác định một tập con các ảnh từ trong tập dữ liệu đại diện cho khu vực đang xét, dựa trên giả định là những ảnh được chụp gần nhau sẽ cùng hiển thị cùng một cảnh. Do tầm quan trọng của tác vụ truy xuất ảnh trong lĩnh vực định vị trực quan mà bài toán nhận diện địa điểm trực quan đã được sinh ra và ngày càng phát triển [7][18][2].

## 1.2 Mục tiêu đề tài

Qua việc phân tích nhu cầu và xác định được những nhóm phương pháp chính đã được ứng dụng trong bài toán định vị trực quan, đề tài đã được xác định sẽ hướng đến các mục tiêu chính:

- *Thứ nhất*, khảo sát những giải pháp đã được đề xuất và phân tích ưu và nhược điểm để có thể chọn ra hướng tiếp cận phù hợp.
- *Thứ hai*, tiến hành kiểm thử kết quả của hướng tiếp cận được chọn làm cơ sở
- *Thứ ba*, thiết kế giải pháp cải thiện dựa trên cơ sở ban đầu.
- *Thứ tư*, đánh giá và kiểm thử giải pháp được đề xuất trên những tập dữ liệu phổ biến, có phạm vi đa dạng và đồng thời tiến hành thực nghiệm trên những thành phần của giải pháp.

Trong giai đoạn *Đồ án chuyên ngành*, nhóm hướng tới hai mục tiêu đầu tiên trong số những mục tiêu đã được trình bày. Hai mục tiêu sau sẽ được thực hiện trong giai đoạn *Đồ án tốt nghiệp*.

### 1.2.1 Khảo sát và phân tích những giải pháp đã có

Nhóm đã tiến hành khảo sát những hướng tiếp cận đã được đề xuất nhằm giải quyết bài toán định vị trực quan trong thời gian gần đây. Thông qua việc khảo sát, nhóm xác định những khía cạnh có thể được cải thiện để đóng góp cho quá trình nghiên cứu của bài toán này.

- Tìm hiểu về những nhóm phương pháp đã được đề xuất
- Tìm hiểu về tác vụ truy xuất ảnh
- Tìm hiểu về những phương pháp ước tính vị trí máy ảnh

Qua quá trình tìm hiểu, nhóm đã quyết định tiếp cận bài toán theo hướng hồi quy vị trí tương đối, xây dựng mô hình gồm hai thành phần được đề xuất trong hai bài nghiên cứu là MixVPR [2] và mô hình hồi quy 2D-2D được đề xuất trong [4] để có thể đạt được kết quả vị trí 6DoF.

### **1.2.2 Tiến hành kiểm thử kết quả của phương pháp cơ sở**

Phương pháp được dùng làm cơ sở cải thiện sẽ được chọn. Những số liệu về hiệu quả của mô hình sẽ được xác nhận lại một lần nữa qua thực nghiệm. Ngoài ra, do phương pháp được nhóm đề xuất sẽ bao gồm hai thành phần khác nhau, nên việc xác định kết quả cơ sở là cần thiết.

- Mô phỏng lại quá trình thí nghiệm trên từng mô hình nhằm tái tạo số liệu được đưa ra trong bài nghiên cứu, xác nhận tính khả thi và hợp lệ.
- Thiết kế một mô hình kết hợp hai thành phần lại và tiến hành thực nghiệm nhằm xác định kết quả cơ sở ban đầu.

### **1.2.3 Thiết kế giải pháp cải thiện hiệu quả của mô hình kết hợp**

Dựa trên mô hình đã được đề xuất, nhóm sẽ tiến hành cải thiện mô hình nhằm giải quyết những vấn đề hiện hữu trong từng thành phần của mô hình

- Đối với thành phần truy xuất ảnh, thể hiện được rằng mô hình có thể cho ra kết quả có độ chính xác cao hơn việc chỉ sử dụng nhãn của ảnh truy xuất được làm kết quả cuối cùng.
- Đối với thành phần hồi quy vị trí, thể hiện được rằng mô hình có thể được áp dụng cho những tập dữ liệu có phạm vi rộng, độ phân bố ảnh thưa hơn.

### **1.2.4 Đánh giá và kiểm thử mô hình**

Để có thể có một cái nhìn khách quan về hiệu quả của mô hình trong bài toán định vị trực quan, thực nghiệm trên những tập dữ liệu có phạm vi và độ phân bố khác nhau sẽ được thực hiện. Cụ thể là

- Thực nghiệm so sánh với những phương pháp thành phần
- Thực nghiệm so sánh với những phương pháp SOTA hiện tại
- Thực nghiệm trên những tập dữ liệu đa dạng
  - Tập dữ liệu có phạm vi nhỏ nhưng phân bố dày đặc: Cambridge Landmarks
  - Tập dữ liệu có phạm vi lớn nhưng phân bố thưa: Pittsburgh 250k
- Thực nghiệm trên những biến thể về nhượng bộ phận của mô hình

## **1.3 Phạm vi đề tài**

- **Mục tiêu chung**

– Hướng đến việc thiết kế và áp dụng thành công sự kết hợp giữa hai mô hình học sâu vào bài toán định vị trực quan. Từ đó cải thiện được điểm yếu của hai hướng tiếp cận.

- **Kết quả mong đợi**

- Đối với truy xuất ảnh, nhóm mong có thể giúp cải thiện độ chính xác của những mô hình truy xuất ảnh đã được đề xuất trước đây.
- Đối với mô hình hồi quy vị trí máy ảnh, nhóm mong có thể giúp mô hình được sử dụng trên các tập dữ liệu lớn hơn.

- **Thời gian**

- Đề án chuyên ngành kéo dài trong 15 tuần
- Đề án tốt nghiệp kéo dài trong 15 tuần

- **Lĩnh vực hướng đến**

- Đề tài hướng đến việc đóng góp thêm một giải pháp mới nhằm góp phần giải quyết được bài toán định vị trực quan, nhằm ứng dụng vào những công nghệ tương tác với thế giới thực.

## 1.4 Cấu trúc đồ án chuyên ngành

Đồ án chuyên ngành sẽ bao gồm năm chương, bao gồm cả chương này. Mỗi chương sẽ bao gồm những nội dung như sau:

- **Chương 1: Giới thiệu**

Trình bày sơ lược về động cơ nghiên cứu, mục tiêu và phạm vi đề tài giải quyết.

- **Chương 2: Các công trình liên quan**

Chương này đề cập tới những hướng đi đã được đề xuất trong các công trình nghiên cứu nhằm giải quyết bài toán định vị trực quan. Ý tưởng và ưu, nhược điểm của mỗi phương pháp sẽ được phân tích nhằm xác định hướng phát triển.

- **Chương 3: Phương pháp đề xuất**

Chương này đề cập đến phương pháp giải quyết bài toán mà nhóm đề xuất bao gồm tổng quan về cơ chế cũng như lý thuyết cách hoạt động.

- **Chương 4: Đánh giá**

Chương này đề cập kết quả khảo sát của nhóm bao gồm kết quả đánh giá hiệu quả của các phương pháp truy xuất ảnh và hồi quy vị trí máy ảnh.

- **Chương 5: Kế hoạch tương lai**

Trình bày tổng quan về quá trình thực hiện và kết quả của giai đoạn *Đồ án chuyên ngành* và đưa ra kế hoạch cho giai đoạn *Đồ án tốt nghiệp*

Cấu trúc của toàn đề tài sẽ được trình bày trong giai đoạn *Đồ án tốt nghiệp*.

## 1.5 Kết chương

Việc xác định chính xác vị trí có một phạm vi ứng dụng rộng rãi, là công nghệ chủ chốt trong rất nhiều lĩnh vực khác nhau. Tuy nhiên, việc chỉ dựa vào hệ thống không đảm bảo được sự ổn định như GPS sẽ hạn chế khả năng phát triển trong tương lai của những lĩnh vực ấy. Vậy nên bài toán định vị hóa trực quan đã được đưa ra để có một hệ thống đưa ra định vị chính xác và ổn định, dựa vào thông tin hình ảnh môi trường xung quanh.

Trong những phương pháp được đưa ra, nhóm tập trung vào việc kết hợp 2 hướng xử lý là truy xuất ảnh và hồi quy tương đối vị trí. Mỗi phương án sẽ có một vấn đề

riêng. Đối với việc truy xuất ảnh, kết quả cho ra được sẽ không có độ chính xác cao, do vị trí của ảnh truy xuất được sẽ được lấy làm kết quả. Đối với hướng hồi quy tương đối vị trí, đa số những phương pháp trước đây đều tập trung vào việc hồi quy trong những không gian nhỏ, có lượng dữ liệu dày đặc, không phù hợp với tập dữ liệu đô thị, mục tiêu của bài nghiên cứu của nhóm. Qua việc ứng dụng cả 2 cách giải quyết trong một mô hình, nhóm hy vọng 2 mô hình có thể bổ trợ, giải quyết điểm yếu của nhau.

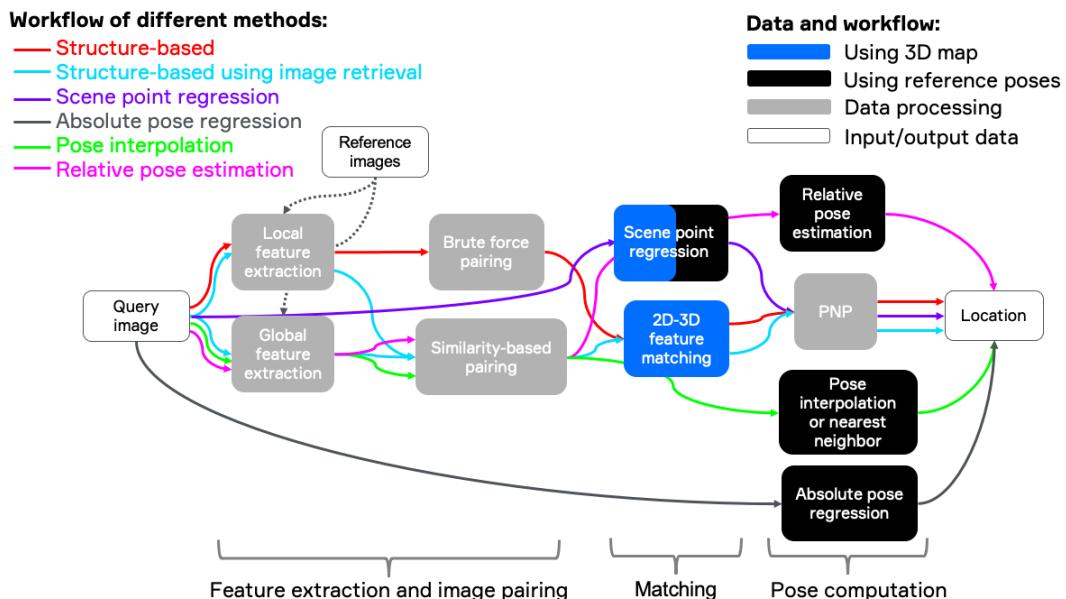
Để có thể lựa chọn được những giải pháp phù hợp, nhóm đã tiến hành khảo sát những kiến thức nền tảng và những công trình nghiên cứu đã được xuất bản trước đây. Những nội dung này sẽ được thể hiện trong **Chương 2: Các công trình liên quan**.

# Chương 2

## CÁC CÔNG TRÌNH LIÊN QUAN

### 2.1 Những phương pháp đã được sử dụng

Định vị trực quan được định nghĩa là một bài toán nhằm xác định được vị trí của máy ảnh từ ảnh chụp được. Cho đến hiện tại, nhiều hướng tiếp cận cho bài toán định vị trực quan đã được đề xuất. Những phương pháp quan trọng nhất có thể được nhóm vào hai nhóm chính là **Những phương pháp sử dụng biểu diễn 3D** và **Những phương pháp hồi quy vị trí**.



Hình 2.1: Tổng quát về những phương pháp định vị trực quan quan trọng [16]

#### 2.1.1 Những phương pháp sử dụng biểu diễn 3D

##### Ý tưởng

Những phương pháp sử dụng biểu diễn 3D sẽ hoạt động xoay quanh việc biểu diễn lại môi trường đang xét bằng một bản đồ đám mây điểm 3D. Bản đồ 3D chứa những đặc trưng trích xuất được từ các ảnh và tiến hành kiểm các cặp đặc trưng tương quan giữa các ảnh với nhau nhằm tạo thành những điểm mô tả 3D trong không gian ba chiều. Những bản đồ này thường sẽ được mô hình tạo ra vào lúc tiền xử lý tập dữ liệu và sẽ được dùng trong suốt quá trình bản địa hóa của mô hình.

## Những bước thực hiện

- Ở bước tiền xử lý, không gian được thể hiện bởi tập dữ liệu sẽ được tái tạo lại trong không gian 3D, sử dụng những ảnh cùng thể hiện một cảnh, nhưng từ những vị trí và góc nhìn khác nhau. Phương pháp tái tạo này được gọi là Tái tạo kiến trúc từ chuyển động - Structure from Motion. Cụ thể hơn, quá trình này sẽ gồm các giai đoạn:
  - Trích xuất mô tả của các đặc trưng từ ảnh: Mỗi ảnh trong tập dữ liệu sẽ được trích xuất đặc trưng và mỗi đặc trưng sẽ được gán một giá trị mô tả để SfM có thể tìm được những đặc trưng tương đồng giữa các hình. Kết quả đầu ra sẽ là một tập các mô tả cho đặc trưng của từng ảnh.
  - Tìm kiếm sự tương quan của các đặc trưng giữa các ảnh: SfM sẽ tìm kiếm những ảnh cùng nhìn bộ phận của cảnh bằng cách kiểm tra sự tương đồng giữa các mô tả của đặc trưng trên các ảnh. Những cặp ảnh chứa những cặp đặc trưng tương quan sẽ cần được xác nhận lại về tính đúng đắn về hình học qua việc có tồn tại một cách biến đổi có khả năng ánh xạ một số lượng trưng giữa hai ảnh.
  - Tái tạo lại cấu trúc: Quá trình tái tạo lại cấu trúc sẽ bắt đầu từ một cặp ảnh khởi tạo. Sau đó, không gian điểm 3D sẽ dần được mở rộng khi các ảnh được thêm tuần tự qua giải thuật PnP trong vòng lặp RANSAC.
- Sau khi có được bản đồ đám mây điểm 3D biểu diễn môi trường đang xét, khi mô hình nhận vào một ảnh mới, những đặc trưng trong ảnh sẽ được trích xuất và so sánh mô tả của chúng với những điểm trong bản đồ 3D của khu vực. Khi những cặp đặc trưng 2D-3D đã được xác định, giải thuật PnP trong vòng lặp RANSAC sẽ được sử dụng một lần nữa để xác định vị trí của ảnh được chụp.

## Những biến thể của phương pháp

Một trong những tác vụ quan trọng nhất trong bài toán định vị bằng biểu diễn 3D chính là việc trích xuất ra được đặc trưng của ảnh và tạo cách mô tả thích hợp cho chúng. Ở những phương pháp truyền thống, những giải thuật được định nghĩa thủ công như SIFT [24] và SURF [6] sẽ được sử dụng. Tuy nhiên, ở những giải thuật này tồn tại một số điểm yếu, xuất phát từ việc chúng không được thiết kế đặc biệt cho tác vụ hiện tại là truy xuất những ảnh biểu diễn cảnh giống nhau. Điều này sẽ làm giảm hiệu quả, đặc biệt trong điều kiện thay đổi như ngày-kém, thay đổi theo mùa. Trong khoảng thời gian gần đây, những hướng tiếp cận theo hướng học sâu đã được đề xuất, chủ yếu sử dụng mạng CNN để trích xuất và biểu diễn đặc trưng. Những cách tiếp cận này vẫn gặp khó khăn trong việc nhận biết được những đặc điểm hình học trong hình [50].

Để có thể biểu diễn chính xác được một thành phố lớn trong không gian 3D, một lượng lớn điểm 3D sẽ cần được sử dụng. Với tập dữ liệu Aachen Day-Night [35], mô hình 3D sẽ có số lượng điểm dao động từ 700 nghìn cho đến 2.5 triệu, tùy thuộc vào số lượng cặp ảnh được dùng. Vì vậy, chiến lược tìm kiếm Brute-Force sẽ dần trở nên bất khả thi khi môi trường biểu diễn trở nên lớn hơn. Để giới hạn lại không gian tìm kiếm, chiến lược truy xuất ảnh đã được sử dụng để tìm kiếm ảnh trong tập dữ liệu tương đồng nhất với ảnh truy vấn. Sau đó, việc tìm kiếm tương quan sẽ chỉ diễn ra trong không gian được thể hiện bởi những ảnh đó [32].

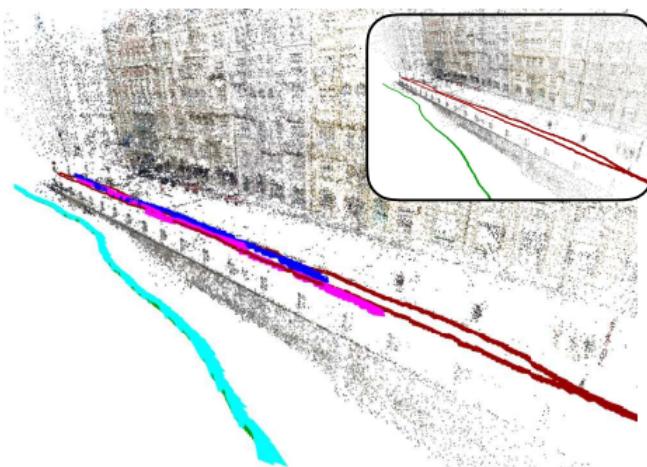
Để không phải lưu trữ một bản đồ 3D cho toàn khu vực đang xét, một phương pháp khác đã được đề xuất là xây dựng bản đồ điểm 3D cục bộ. Thay vì sinh ra bản đồ 3D cho toàn bộ khu vực đang xét ở bước tiền xử lý, phương pháp này sẽ sử dụng những ảnh truy xuất được để sinh ra bản đồ cục bộ [36]. Phương pháp này có thể tận dụng được những điểm mạnh của việc sử dụng cách biểu diễn 3D mà không cần duy

trí bản đồ điểm 3D của toàn bộ khu vực. Tuy nhiên, cách tiếp cận này vẫn có một số điểm yếu như việc thời gian chạy trở nên đáng kể và số lượng ảnh truy xuất được có thể chưa đủ để sinh ra một bản đồ 3D phù hợp.

Một cách tiếp cận khác với mục tiêu loại bỏ sự cần thiết của bản đồ 3D, sử dụng mạng học sâu để xác định được vị trí của một điểm ảnh trong môi trường 3D, gọi là hồi quy điểm trong cảnh - Scene Point Regression [8]. Tuy nhiên, những mô hình thuộc phương pháp này không có tính khái quát hóa mà để có thể chạy trên một khu vực nhất định thì mô hình cần dữ liệu mẫu thuộc khu vực đó để huấn luyện.

### **Phân tích ưu và nhược điểm của phương pháp**

Nhờ vào việc sử dụng cách biểu diễn trong không gian 3D, lớp phương pháp này đã thành công học được tính hình học của môi trường, nhờ vậy nên những kết quả của lớp phương pháp này thường sẽ đạt được kết quả định vị chính xác. Ngoài ra, lớp phương pháp này sử dụng điểm 3D trong không gian làm mốc để định vị. Nhờ vậy nên mô hình vẫn sẽ cho ra được kết quả tốt khi không gian mẫu của vị trí chụp của ảnh bị giới hạn.



Hình 2.2: Kịch bản khi mà tập dữ liệu huấn luyện bị hạn chế. Quỹ đạo của tập huấn luyện và tập kiểm thử có màu **đỏ** và **xanh lá**. Kết quả của những mô hình lấy vị trí của ảnh làm mốc là PoseNet [21] và MapNet [9] có quỹ đạo kết quả màu **xanh dương** và **tím**. Kết quả của phương pháp sử dụng biểu diễn 3D là Active Search [34] có quỹ đạo màu **xanh lam**

Đa số những vấn đề xoay quanh lớp phương pháp này sẽ liên quan đến việc sinh ra, lưu trữ và sử dụng biểu diễn 3D của khu vực. Do những bản đồ 3D biểu diễn những khu vực lớn sẽ có nhiều điểm 3D, dẫn đến việc tiêu tốn nhiều tài nguyên tính toán để sinh ra và sử dụng, đồng thời cần nhiều bộ nhớ để lưu trữ. Ngoài ra, đánh đổi cho khả năng biểu diễn một khu vực với độ chi tiết cao, những mô hình sử dụng cách biểu diễn 3D sẽ mất khả năng khái quát hóa khi xử lý những ảnh bên ngoài khu vực đang xét và cần phải được huấn luyện lại trên tập dữ liệu khác.

#### **2.1.2 Những phương pháp hồi quy vị trí**

##### **Ý tưởng**

Với những thành công gần đây trong các tác vụ của ngành khoa học máy tính như phân loại ảnh, phân vùng ảnh theo ngữ nghĩa, truy xuất ảnh, những phương pháp sử

dụng mạng học sâu đã có thể phần biểu diễn được thông tin về hình học bên trong mô hình. Vậy nên, những phương pháp học sâu, sử dụng CNN dần được phổ biến trong bài toán định vị trực quan. Các thành phần của một giải pháp có thể được thay thế hoàn toàn, hoặc thay thế cụ thể từng bộ phận.

### Những bước thực hiện

Đầu tiên, về hướng tiếp cận áp dụng phương pháp học sâu đầu cuối, mô hình sẽ có thể nhận vào một ảnh RGB và xác định được vị trí chụp của ảnh đó mà không cần có sự can thiệp của con người. Thông thường, những mô hình này sẽ bao gồm những bước xử lý ảnh như sau:

- Mã hóa ảnh: mạng CNN sẽ được dùng ở đầu vào để có thể trích xuất được những chi tiết trong ảnh, sử dụng để sinh ra một giá trị mã hóa, đại diện cho ảnh. Ở bước này, những kiến trúc mạng CNN khác nhau như VGG, ResNet, EfficientNet có thể được thử nghiệm để tìm ra mạng thích hợp nhất với tác vụ định vị trực quan. Với những thành công gần đây trên những tác vụ thị giác máy tính khác, Vision Transformer có thể giúp cải thiện quá trình mã hóa.
- Ánh xạ đến giá trị vị trí của ảnh: Từ những giá trị mã hóa đại diện cho ảnh, mô hình học sâu sẽ ánh xạ đến vị trí thực tế của ảnh, thường ở dạng 6DoF. Như trong PoseNet [21], các lớp kết nối đầy đủ sẽ được sử dụng như hàm ánh xạ ra vị trí và hướng quay của ảnh.

### Những biến thể của phương pháp

#### Phân tích ưu và nhược điểm của phương pháp

##### 2.1.3 Những phương pháp khác

##### 2.1.4 Xác định những vấn đề hiện hữu trong bài toán định vị trực quan

###### • Những phương pháp hồi quy vị trí

- Vấn đề về phạm vi hoạt động của mô hình: Đa số những mô hình hồi quy vị trí đều chỉ hoạt động tốt trên những tập dữ liệu có phạm vi nhỏ và phân bố dày đặc như 7Scenes [40] và Cambridge Landmarks [21]. Đối với những tập dữ liệu có phạm vi rộng và phân bố thưa như Aachen Day-Night [37], chỉ phương pháp nội suy đơn giản đã được áp dụng và kết quả đem lại là không quá tốt.
- Vấn đề về khả năng khái quát hóa: Những mô hình hồi quy tuyệt đối sẽ xây dựng ngầm biểu diễn của khu vực đang xét bên trong những trọng số của mô hình. Vì vậy nên, một mô hình được huấn luyện trên một không gian nhất định chỉ có thể xử lý những ảnh được chụp trong không gian đó. Đối với những mô hình hồi quy tương đối, vấn đề này sẽ phụ thuộc vào cơ chế truy xuất ảnh được sử dụng.
- Vấn đề về độ chính xác: Những phương pháp hồi quy sẽ không cho ra kết quả chính xác như những phương pháp sử dụng cách biểu diễn 3D.

###### • Những phương pháp truy xuất ảnh

- Vấn đề về tài nguyên xử lý: Tuy bài toán VPR đã có nhiều bước đột phá trong thời gian qua, đa số các phương pháp đạt kết quả tốt đều sử dụng những cách tổng hợp mô tả của các đặc trưng trong ảnh như NetVLAD [3] hoặc những biến thể tích hợp cơ chế tập trung [18], ngữ nghĩa của bộ phận ảnh [29], bối cảnh của ảnh [17]. Những cơ chế đó sẽ có ảnh hưởng tiêu cực đến thời gian chạy do độ phức tạp của giải thuật cũng như độ lớn của không gian mô tả.

- Vấn đề về khả năng biểu diễn của CNN: Những cách tiếp cận theo mô hình học sâu sử dụng CNN cho đến hiện tại thường sẽ gặp những vấn đề trong việc biểu diễn được những thông tin về môi trường địa lý trong bài toán định vị trực quan.

## 2.2 Nhận dạng địa điểm trực quan - Visual Place Recognition

Trước đây, việc định vị trực quan trên quy mô lớn (large-scale visual localization) được coi là một vấn đề truy xuất hình ảnh[48]. Vị trí cho hình ảnh truy vấn được xác định bởi hình ảnh tương tự nhất được lấy từ cơ sở dữ liệu. Tuy nhiên, để đáp ứng nhu cầu xác định vị trí của ảnh với độ chính xác trên 6 bậc tự do (6 Degrees of Freedom), việc sử dụng mô hình 3D để ước tính tư thế của máy ảnh ngày càng được các nhà nghiên cứu đề xuất và sử dụng. Sử dụng phương pháp này, bài toán định vị trực quan có thể được chia thành hai bước: truy xuất hình ảnh và định vị tư thế máy ảnh từ hình ảnh truy xuất được.

### 2.2.1 Học biểu diễn - Representational learning

### 2.2.2 Học biểu diễn NetVLAD

### 2.2.3 Tối ưu hóa đặc trưng - MAC

### 2.2.4 Trung bình Hölder - Trung bình GeM

### 2.2.5 Truy xuất và tái xếp hạng

### 2.2.6 Học biểu diễn bằng Vision Transformer

### 2.2.7 Học biểu diễn bằng Feature Mixer

## 2.3 Ước tính vị trí của máy ảnh - Pose Estimation

Ước tính vị trí máy ảnh (Camera Pose Estimation) là một bài toán thuộc chuyên ngành thị giác máy tính nhằm xác định vị trí (position) và góc quay (orientation) chính xác nhất có thể của máy ảnh thông qua dữ liệu hình ảnh được chụp từ chính máy ảnh. Đây là một bước cực kỳ quan trọng trong việc giải quyết bài toán định vị trực quan, thường được áp dụng sau khi bước nhận dạng địa điểm trực quan đã trích xuất được ảnh từ kho dữ liệu. Hiện nay, có tương đối nhiều hướng tiếp cận đối với bài toán này. Một trong những phương pháp phổ biến nhất là huấn luyện một mô hình học sâu để xác định 6 chiều tự do (Degree of Freedom) từ số ít ảnh (Absolute Pose Regression và Relative Pose Regression) hoặc xây dựng một mô hình 3D từ tập dữ liệu có sẵn rồi tiến hành chuyển ảnh đầu vào sang các điểm 3D để dễ dàng so sánh và xác định vị trí (Structure From Motion).

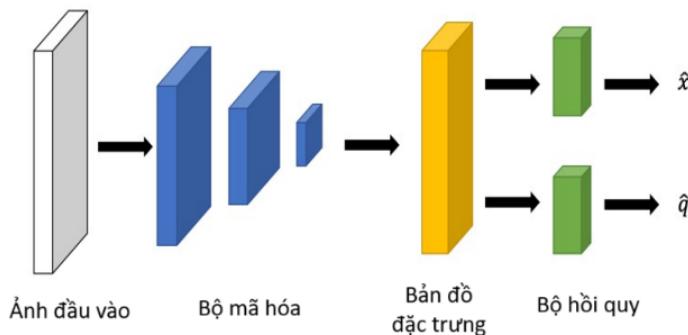
### 2.3.1 Hồi quy vị trí tuyệt đối - Absolute Pose Regression

Hồi quy vị trí tuyệt đối (Absolute Pose Regression) hướng đến việc dự đoán vị trí và góc quay chính xác nhất của ảnh bằng một mô hình mạng nơ-ron tích chập thông qua việc cải thiện trọng số của mô hình. Tùy thuộc vào đâu vào của mô hình mà hồi

quy vị trí tuyệt đối được chia thành ba hướng chính: hồi quy vị trí tuyệt đối với một ảnh, chuỗi ảnh hoặc đoạn phim.

### Hồi quy vị trí tuyệt đối đơn ảnh - Absolute pose regression through single monocular image

Với phương pháp hồi quy vị trí tuyệt đối thông qua một ảnh (Absolute pose regression through single monocular image), quy trình chung thường bao gồm: đầu vào - mạng - đầu ra. Đầu vào sẽ là một ảnh RGB với đầu ra là vị trí 6 độ tự do của máy ảnh. Thông thường, kiến trúc của mạng lưới tính toán sẽ bao gồm các thành phần như sau: bộ mã hóa, bộ định vị, bộ hồi quy.



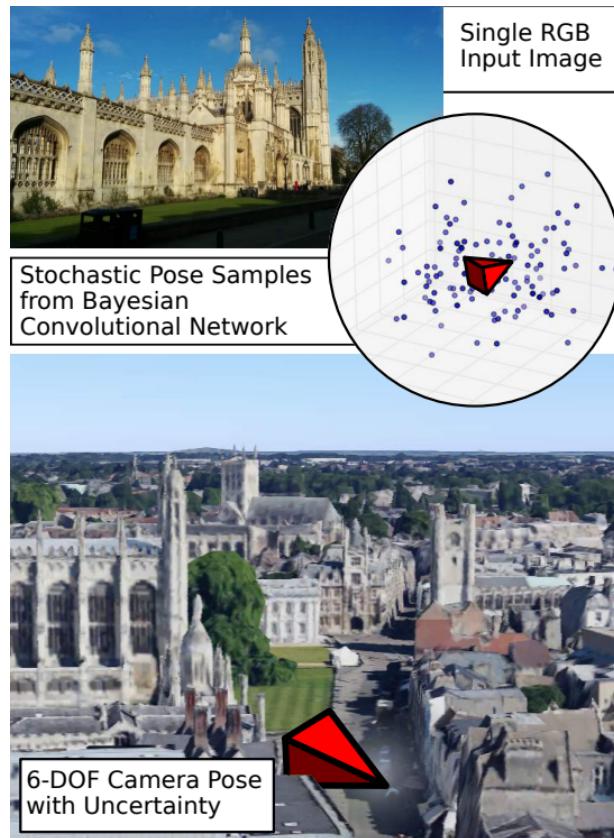
Hình 2.3: Kiến trúc mô hình hồi quy vị trí tuyệt đối đơn ảnh [21]

#### Phương pháp sử dụng hàm mất mát Euclidean cố định:

PoseNet [21] là công trình đầu tiên huấn luyện mô hình mạng nơ-ron tích chập để hồi quy vị trí máy ảnh từ một ảnh RGB, hoàn toàn không phụ thuộc vào bất kỳ cơ chế bên ngoài nào khác. Vào thời điểm ra mắt, PoseNet đã cho thấy sự vững chắc của mô hình vượt trội so với phương pháp tái tạo kiến trúc từ chuyển động dựa trên cơ chế "biến đổi tính năng bất biến tỷ lệ" (Scale-invariant Feature Transform Structure from Motion): độ hiệu quả của kiến trúc về sau giảm mạnh nếu độ lớn của tập dữ liệu huấn luyện giảm đến một mức nhất định. Hàm mất mát Euclidean của PoseNet được định nghĩa như sau:

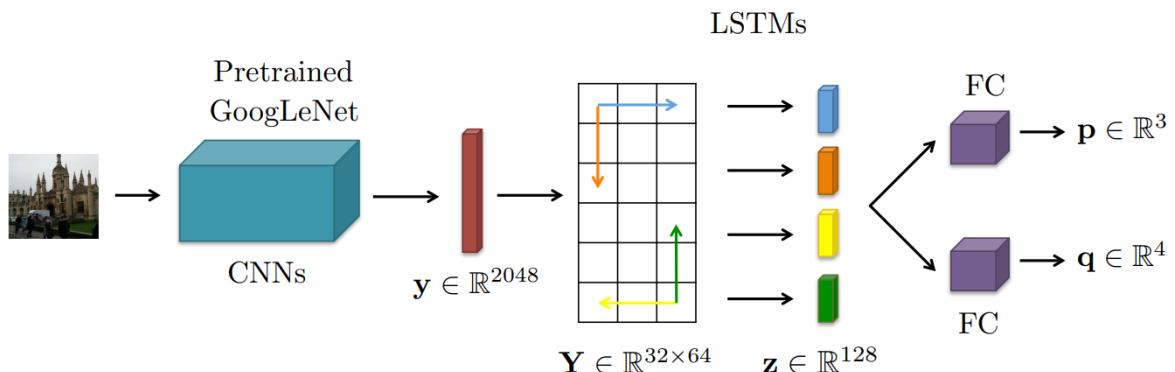
$$loss(I) = \|\hat{x} - x\|_2 + \beta \left\| \hat{q} - \frac{q}{\|q\|} \right\|_2$$

Kế thừa từ PoseNet, đã có nhiều công trình và bài báo tìm cách cải thiện phương pháp định vị hoặc thay thế hàm mất mát nhằm nâng cao hiệu suất chung của toàn kiến trúc. Với các công trình có mong muốn cải thiện hàm mất mát của mô hình, một chiến thuật chung là kết hợp hàm mất mát Euclidean và phương pháp giảm độ dốc Stochastic. Về mặt cải thiện hiệu quả định vị cũng như tìm hiểu về độ thiếu chính xác của mô hình, một nhóm tác giả đề xuất thêm xác suất Bernoulli vào mô hình nơ-ron tích chập [19] nhằm xác định độ thiếu chính xác của mô hình. Ý tưởng chính của phương pháp này là xác định và tận dụng độ thiếu chính xác để dự đoán sai số trong định vị, phương pháp này đã cải thiện độ chính xác cho PoseNet cho cả những cảnh ngoài trời và bên trong nhà.



Hình 2.4: Minh họa mô hình CNN được áp dụng phân phối Bernoulli [19]

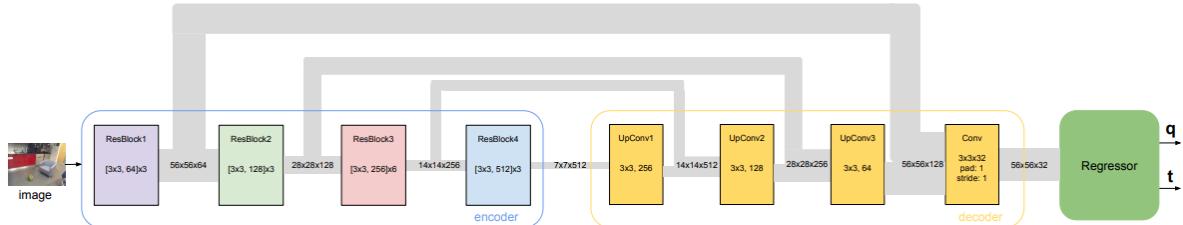
Từ những thông tin được nêu ra trong bài báo [21], ta biết được rằng mô hình PoseNet có một lớp kết nối đầy đủ với 2048 chiều, tạo điều kiện cho việc áp dụng một lớp bộ nhớ dài ngắn hạn để giảm chiều đặc trưng giúp cải thiện độ chính xác định vị [45, 46]. Nhóm tác giả Watch và cộng sự [45] đề xuất tận dụng các lớp bộ nhớ dài ngắn hạn lên đầu ra của PoseNet để giảm chiều và chọn ra những đặc trưng hữu ích nhất cho bài toán định vị vị trí. Các thí nghiệm đo lường cho thấy phương pháp này vượt trội hơn PoseNet khoảng 30% về sai số vị trí và 55% về sai lệch góc quay.



Hình 2.5: Minh họa kiến trúc mô hình LSTM PoseNet [45]

Để cải thiện độ chính xác định vị, một kiến trúc đồng hồ cát được đề xuất với việc thêm một phần chức năng mã hóa thông tin hữu ích từ kiến trúc vật thể thô và một phần chức năng thu chi tiết vật thể mịn. Hourglass PoseNet [25] có kiến trúc gồm ba thành phần chính là bộ mã hóa, bộ giải mã và bộ hồi quy. Mô hình này sử dụng một

mô hình ResNet34 đã được tinh chỉnh làm bộ mã hóa - giải mã. SVS PoseNet [28] dùng mô hình VGG16 kết hợp thêm hai lớp kết nối dày đủ để có thể dự đoán riêng vị trí và góc quay. BranchNet [47] sử dụng mô hình mạng hai nhánh học đồng thời biểu diễn góc quay và độ dời để giảm thiểu độ thua của các vị trí được lấy mẫu một cách hiệu quả. Dù hướng tiếp cận có sự khác biệt, các công trình trên đều có cùng hàm mất mát với PoseNet.

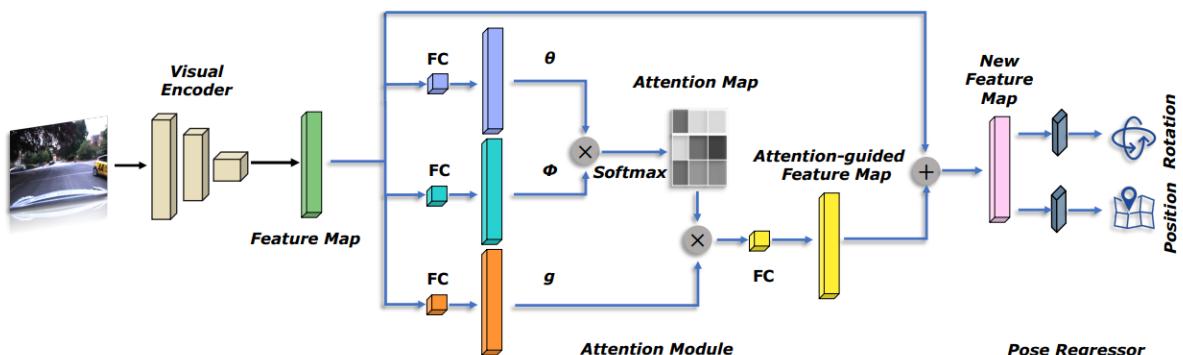


Hình 2.6: Minh họa kiến trúc mô hình Hourglass PoseNet [25]

### Phương pháp sử dụng hàm mất mát có trọng số học được:

Để học được thông tin về vị trí và góc quay từ dữ liệu ảnh, hàm mất mát cố định Euclidean áp dụng các siêu tham số cân bằng để giúp việc học thông tin vị trí và góc quay một cách độc lập, tuy nhiên để học trọng số thì rất tốn kém. Geometric PoseNet [20] đề xuất sử dụng hàm mất mát vị trí có trọng số học được để cân bằng hiệu suất và cải thiện độ ổn định. Khi so sánh với PoseNet, phương pháp này giữ lại độ mở rộng và độ chắc chắn mà không cần chỉnh sửa các siêu tham số cố định cân bằng trong hàm mất mát.

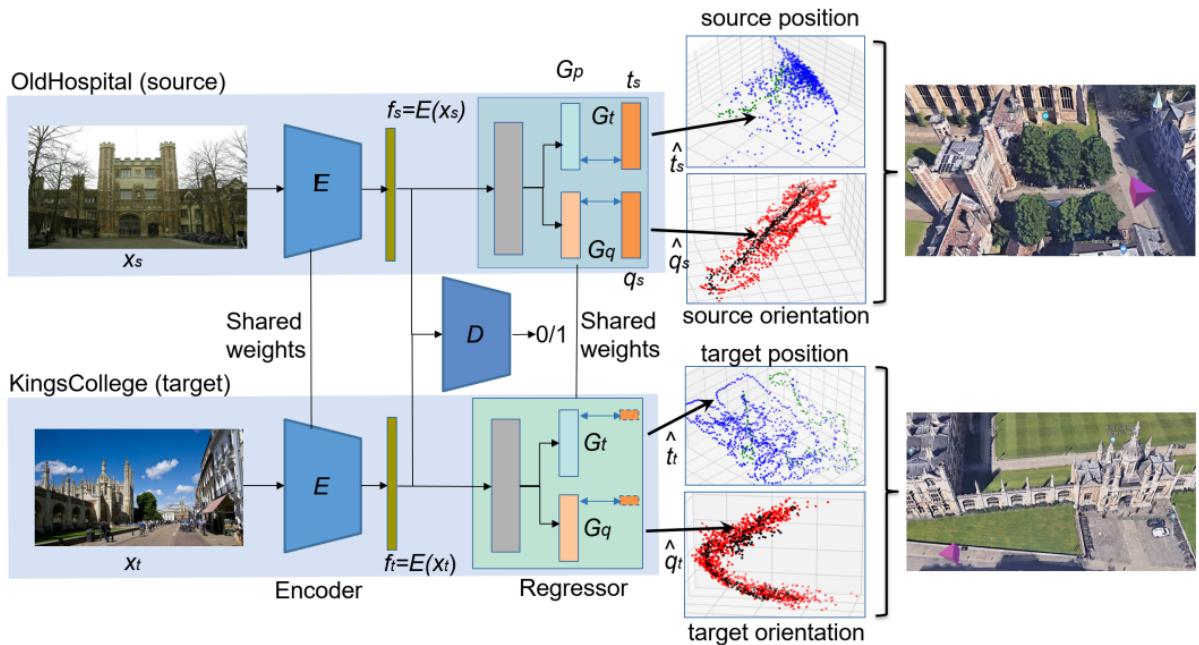
AtLoc [46] thêm vào mô hình một mô-đun tập trung (Attention Module) trước khi xác định các tọa độ hồi quy để ép buộc mạng phải tập trung vào phần chính - phần mang nhiều thông tin hữu ích nhất của hình ảnh đầu vào. Ngoài ra, AtLoc sử dụng ResNet34 được huấn luyện sẵn trên tập dữ liệu ImageNet làm bộ mã hóa, sau đó hồi quy lớp kết nối dày đủ 2048 chiều của PoseNet.



Hình 2.7: Minh họa kiến trúc mô hình AtLoc [46]

AdPR [10] thêm một mạng phân biệt và học đối lập. Điều này không chỉ hồi quy vị trí mà còn tinh chỉnh vị trí. Khi trích xuất đặc trưng, AdPR áp dụng mạng ResNet-18, vì nó có thể đạt được hiệu suất tốt nhất so với VGG16 và AlexNet. APANet [13] cũng sử dụng một mạng đối lập để tạo ra hình ảnh liên quan đến hình ảnh đầu vào để ước lượng tốt hơn vị trí của máy ảnh. Một mô-đun Dropout được thêm trước bộ mã hóa trích xuất đặc trưng để xuất ra nhiều khả năng không chắc chắn, điều này có thể cải thiện độ chắc chắn của mô hình dưới điều kiện thử thách như thay đổi vị trí,

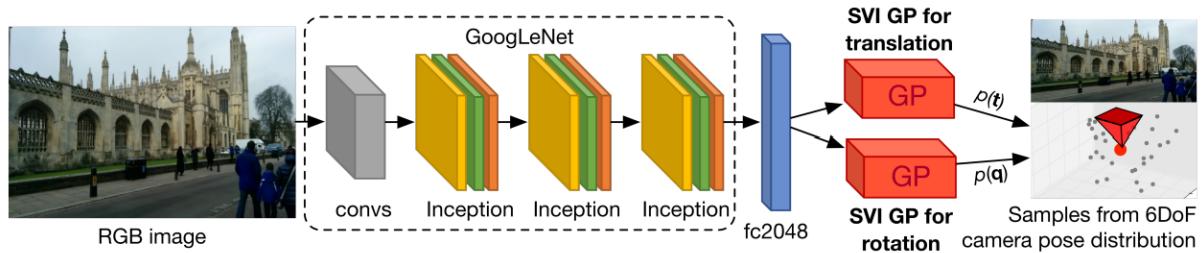
thời tiết,... . Sau khi trích xuất, mô-đun tập trung tự động được thêm để điều chỉnh lại trong số bản đồ đặc trưng.



Hình 2.8: Minh họa kiến trúc mô hình APANet [13]

### Phương pháp sử dụng hàm mất mát khác:

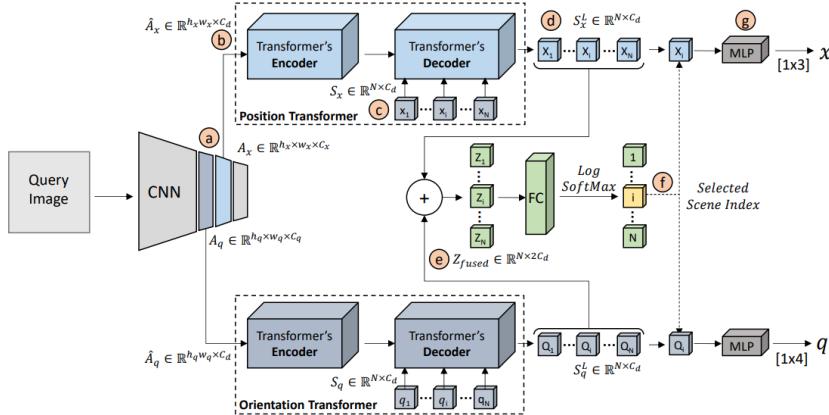
Không dùng đến cả hàm mất mát cố định hoặc những hàm mất mát có trọng số học được, một số nhóm nghiên cứu đề xuất nên cân nhắc sử dụng các mô-đun khác để cải thiện hiệu suất định vị. GeoPoseNet [20] đề xuất sử dụng hàm mất mát tái chiểu: đặc tả sai sót tái chiểu của cảnh. Hàm mất mát tái chiểu chuyển mất mát chung học được thành khác biệt tọa độ ảnh, do đó có thể thay đổi trọng số giữa vị trí và góc quay, tùy thuộc vào các cảnh khác nhau trong quá trình huấn luyện mô hình. GPoseNet [11] xây dựng mô hình mới bằng cách thêm vào hai bộ "Hồi quy quá trình Gaussian suy luận biến phân ngẫu nhiên" (Stochastic Variational Inference Gaussian Process Regressions - SVI GPs) sau lớp kết nối đầy đủ để học phân phối xác suất của vị trí - hướng quay đầu ra và giảm tần suất sử dụng siêu tham số. Hàm mất mát của GPoseNet kết hợp hàm mất mát SVI GPs sử dụng ranh giới điều kiện dưới của hai xác suất tích lũy log  $L_s vi$  và hàm mất mát CNN với siêu tham số  $\beta_{g_t}$  và  $\beta_{n_q}$  của PoseNet.



Hình 2.9: Minh họa kiến trúc mô hình GPoseNet [20]

Một nhóm nghiên cứu [38, 39] đề xuất áp dụng mô hình Transformer vào tác vụ hồi quy vị trí tuyệt đối. Mô hình nhận vào một ảnh đơn và sử dụng một CNN làm bộ trích xuất đặc trưng, sau đó các bản đồ đặc trưng được truyền song song qua hai

nhánh: mỗi nhánh là một mô hình Transformer đảm nhiệm một tác vụ riêng lần lượt là hồi quy vị trí và hồi quy hướng quay. Mô hình sử dụng hàm mất mát tương tự với PoseNet.



Hình 2.10: Minh họa kiến trúc mô hình Multi-Scene Transformer [38, 39]

### Hồi quy vị trí tuyệt đối đa ảnh - Absolute pose regression through auxiliary image sequence

Một phương pháp khác được áp dụng để hồi quy vị trí tuyệt đối là áp dụng học có hỗ trợ với một chuỗi ảnh. Học có hỗ trợ được định nghĩa là phương pháp cải thiện hiệu suất của một tác vụ chính thông qua việc học cùng lúc các tác vụ hỗ trợ. Phương pháp học này giúp mô hình phát triển các biểu diễn dữ liệu tốt hơn. Bằng cách tận dụng các tác vụ hỗ trợ có liên quan khác trong quá trình học, hiệu suất của tác vụ chính có thể được cải thiện. Ở đây, học có hỗ trợ ám chỉ việc kết hợp APR với các tác vụ phụ có liên quan như đo lường cảm biến trực quan. Hàm mất mát của các phương pháp học có hỗ trợ thường bao gồm hàm mất mát của APR kết hợp với hàm mất mát của các phương pháp phụ trợ, thậm chí có thể kết hợp cả hàm mất mát của APR và RPR. Khác với các phương pháp hồi quy vị trí tuyệt đối đơn ảnh, phương pháp học có hỗ trợ học từ các cặp ảnh với bản chất là học cách xác định vị trí tuyệt đối bằng cách đánh giá trước hết vị trí tương đối với các ràng buộc phụ thuộc.

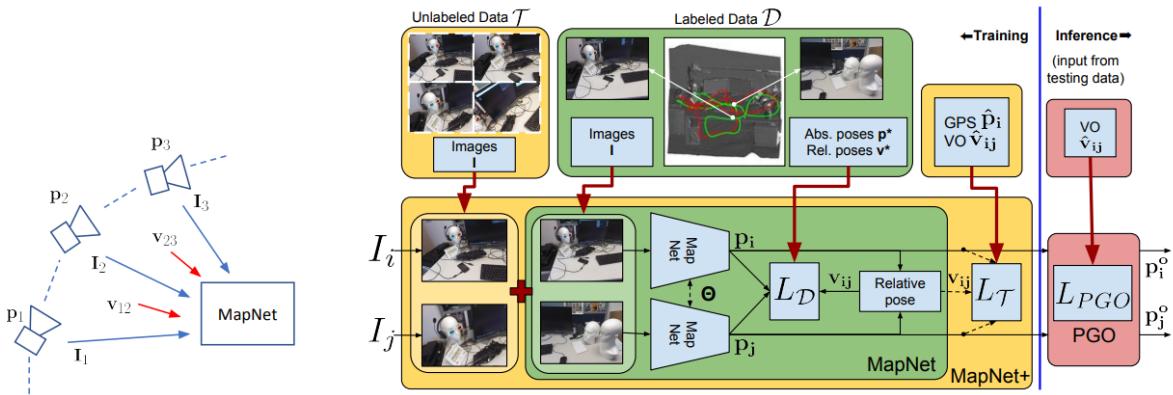
MapNet [9] đề xuất thêm một thuật ngữ mất mát lấy từ các cặp ảnh làm một ràng buộc hình học, điều này đã có thể cải thiện mạnh mẽ hiệu suất khả năng định vị. Về hàm mất mát, MapNet giảm thiểu tối đa cả mất mát vị trí tuyệt đối cho mỗi hình ảnh và mất mát vị trí tương đối giữa các cặp hình ảnh:

$$l(I_{total}) = l(I_i) + \alpha \sum_{i \neq j} loss(I_{ij})$$

Trong đó,  $loss(I_{ij})$  ám chỉ vị trí máy ảnh tương đối  $p_i$  và  $p_j$  giữa các cặp hình ảnh  $I_i$  và  $I_j$ , được tính bởi hàm mất mát với trọng số có thể học được.

Thêm vào đó, MapNet chuyển một giá trị quaternion thành logarit của giá trị đó - biểu diễn phép quay ba độ tự do (3DoF) với ba chiều chưa bị tham số hóa quá mức.  $logq$  được biểu diễn như dưới đây, với  $u$  và  $v$  đại diện cho phần thực và ảo của một quaternion đơn vị:

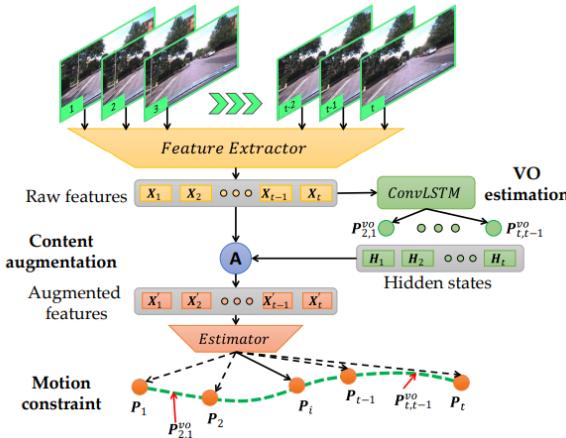
$$logq = \begin{cases} \frac{v}{\|v\|} \cos^{-1} u, & \|v\| \neq 0 \\ 0 & \end{cases}$$



Hình 2.11: Minh họa kiến trúc mô hình MapNet [9]

Năm 2019, tác giả Xue và những cộng sự [49] cũng có một hướng tiếp cận tương tự khi hồi quy vị trí máy ảnh thông qua những ràng buộc về không gian - thời gian, trong đó đặc trưng cục bộ cải thiện định vị toàn cục - gọi là "Cục bộ hỗ trợ toàn cục" (Local Support Global - LSG). Thêm vào đó, LSG đề xuất sử dụng một đánh giá đã được “tăng cường nội dung” để ước lượng lỗi vị trí và tinh chỉnh dựa trên chuyển động, để tối ưu hóa dự đoán vị trí thông qua các ràng buộc chuyển động. LSG sử dụng một hàm mất mát vị trí toàn cầu  $L_g$  lấy từ hồi quy tuyệt đối, hàm mất mát hồi quy đo lường cảm biến trực quan  $L_{vo}$ , các ràng buộc hình học và hàm mất mát liên kết chuyển động  $L_{joint}$  để tối ưu hóa hồi quy vị trí như sau:

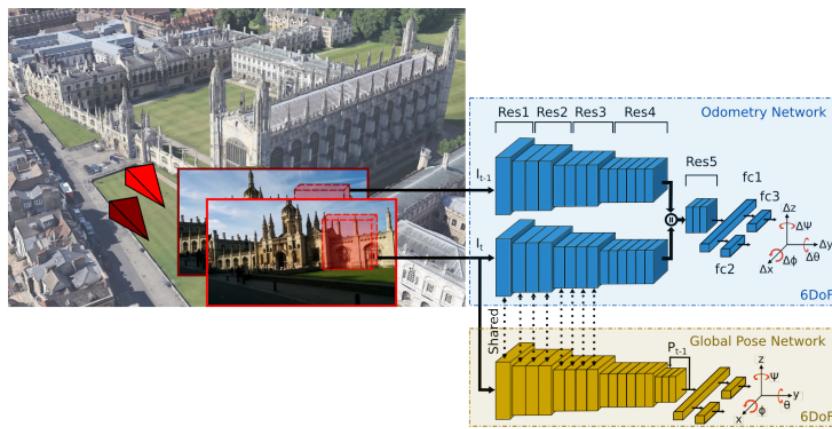
$$l_{total} = l_g + l_{vo} + l_{joint}$$



Hình 2.12: Minh họa kiến trúc mô hình LSG [49]

VlocNet [44] cũng học đồng thời đo lường cảm biến trực quan như một tác vụ phụ để hồi quy vị trí toàn cục với hai mạng phụ. Mất mát nhất quán hình học được điều chỉnh để giảm thiểu tối đa lỗi vị trí, được định nghĩa như sau:

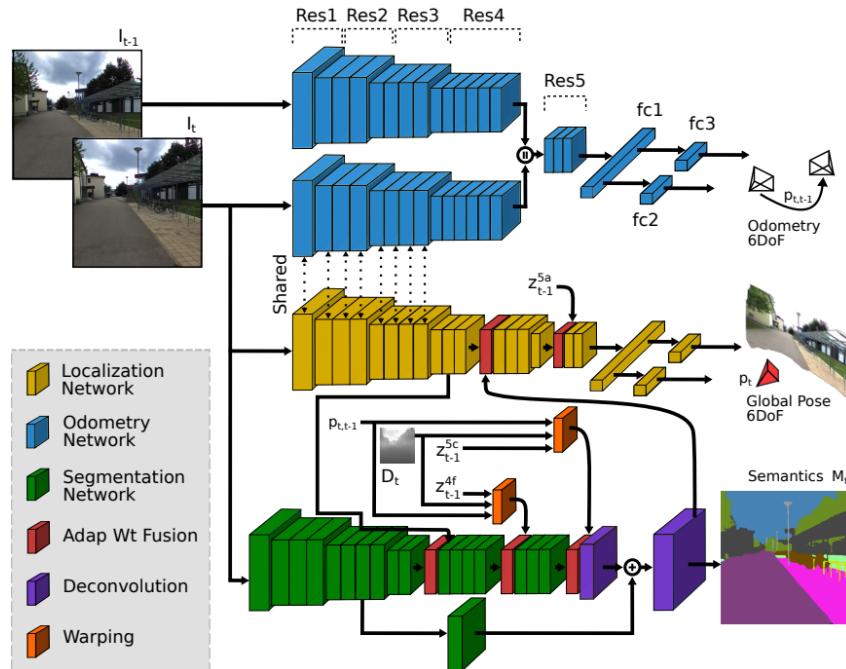
$$l(I_{total}) = (I_{i_x} + I_{ij_x}) \exp(-\hat{s}_x) + (I_{i_q} + I_{ij_q}) \exp(-\hat{s}_q) + \hat{s}_q$$



Hình 2.13: Minh họa kiến trúc mô hình VlocNet [44]

VlocNet++ [30] giới thiệu kiến thức ngữ nghĩa vào hồi quy vị trí, kết hợp thông tin hình học-thời gian với các đặc trưng ngữ nghĩa cùng một lúc. Hàm mất mát của VlocNet++ kết hợp hồi quy vị trí toàn cục, mất mát đo lường cảm biến trực quan và mất mát Entropy chéo cho mất mát phân đoạn ngữ nghĩa cùng một lúc, với ba yếu tố  $\hat{s}_{log}$ ,  $\hat{s}_{vo}$  và  $\hat{s}_{seg}$  để cân bằng ba thành phần này:

$$l(I_{total}) = l_{loc} \exp(-\hat{s}_{loc}) + \hat{s}_{loc} + l_{vo} \exp(-\hat{s}_{vo}) + \hat{s}_{vo} + l_{seg} \exp(-\hat{s}_{seg}) + \hat{s}_{seg}$$



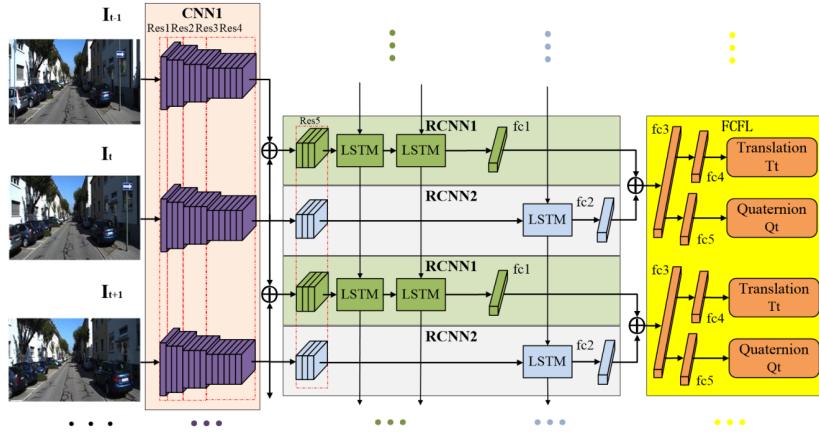
Hình 2.14: Minh họa kiến trúc mô hình VlocNet++ [30]

Một bản mở rộng của AtLoc, AtLocPlus [46] cũng kết hợp các ràng buộc thời gian để học cùng lúc mất mát vị trí tuyệt đối và mất mát vị trí tương đối, dẫn đến hiệu suất tốt hơn AtLoc trong việc sử dụng một đầu vào ảnh đơn. AtLocPlus sử dụng hàm mất mát giống với MapNet.

DGRNet [23] đề xuất một kiến trúc với một mạng con hồi quy vị trí tương đối RCNN1, một mạng con hồi quy vị trí toàn cục RCNN2 và lớp kết hợp kết nối đầy đủ dùng để trích xuất đặc trưng từ ảnh. Ràng buộc biến đổi chéo (Cross transformation constraint – CTC) và sai số toàn phương trung bình (Mean squared error – MSE) được

áp dụng vào hàm mất mát để cải thiện hiệu suất hồi quy. DGRNet đã sử dụng kết hợp hàm mất mát tương đối, toàn cục, CTC  $\hat{l}_i$  và sự thật nền tảng  $\hat{P}_i$  như sau:

$$w = \underset{w}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^N \sum_{k=0}^6 (l_k^i) + \sum_{j=0}^4 \left\| P^i j - \hat{P}_j^i \right\|_2^2$$



Hình 2.15: Minh họa kiến trúc mô hình DGRNet [23]

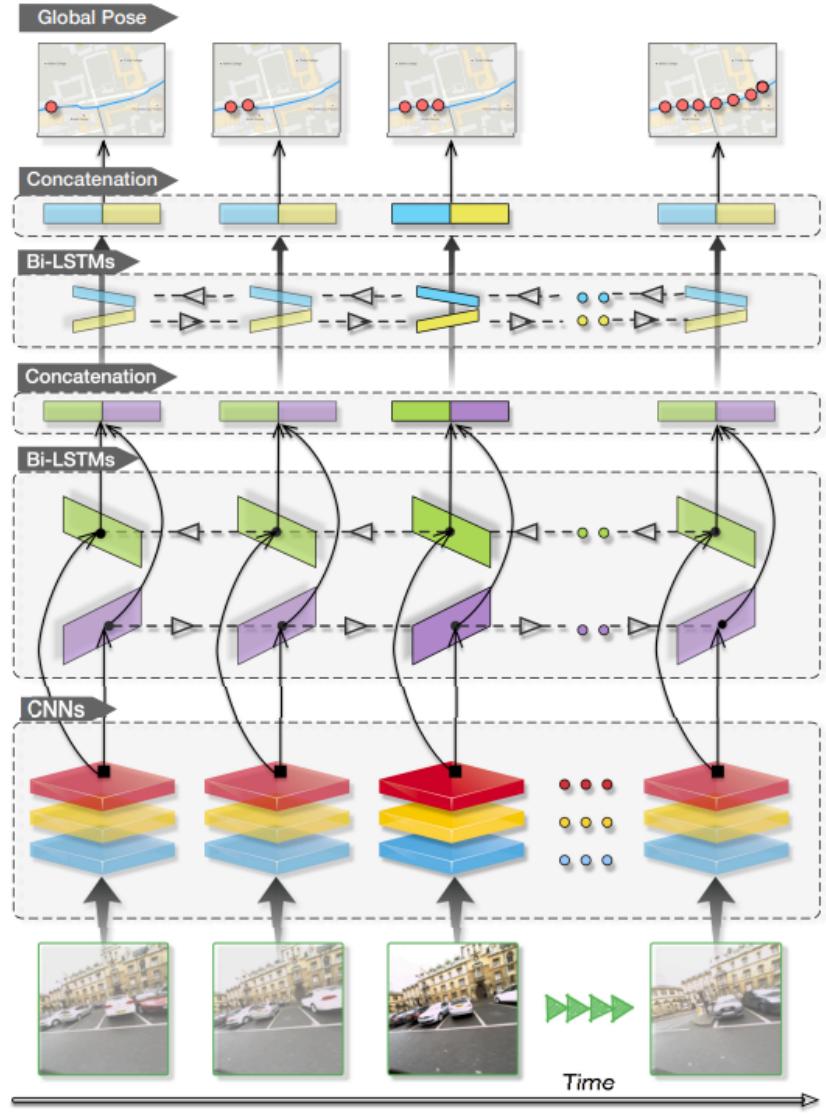
### Hồi quy vị trí tuyệt đối qua đoạn phim - Absolute pose regression through video

Không chỉ đơn ảnh hay đa ảnh, ngay cả đoạn phim cũng có thể được sử dụng để thêm một ràng buộc thời gian có tính trơn tru hơn cho hồi quy vị trí. Các đoạn phim hay các dữ liệu cảm biến khác đều có thể được truy cập dễ dàng bởi các thiết bị di động. Đoạn phim có thể được đồng bộ hóa với các dữ liệu đầu vào khác như đo lường cảm biến trực quan, các cảm biến IMU như đồng hồ tăng tốc và đồng hồ quay và dữ liệu GNSS bằng thông tin thời gian, cụ thể là bằng cách căn chỉnh các mốc thời gian. Với một quy trình tương tự như các phương pháp ARP dựa trên hình ảnh đơn và chuỗi hình ảnh, ARP dựa trên đoạn phim cũng hồi quy độ dời và hướng quay thông qua bộ trích xuất đặc trưng là một mạng nơ-ron tích chập và bộ hồi quy vị trí cục bộ.

VidLoc [14] đề xuất một mô hình hồi quy vị trí máy ảnh dựa trên việc kết hợp CNN – RNN, mục đích là có thể khiến quá trình dự đoán vị trí từ ảnh hay một đoạn phim được trơn tru hơn. Mạng được xây dựng bằng cách sử dụng GoogLeNet Inception [42] mà không dùng đến lớp kết nối đầy đủ để trích xuất đặc trưng ảnh và một mô-đun LSTM hai chiều để mô hình hóa thông tin thời gian với các ô nhớ. Hàm mất mát của VidLoc được tính bằng tổng của lỗi vị trí và lỗi hướng quay từ đầu ra của LSTM như sau:

$$l = \sum_{t=1}^T \alpha_1 \|x_t - \hat{x}_t\| + \alpha_2 \|q_t - \hat{q}_t\|$$

Với  $[x_t, q_t]$  và  $[\hat{x}_t, \hat{q}_t]$  đại diện cho sự thật nền tảng và giá trị dự đoán cho độ dời vị trí và hướng quay.



Hình 2.16: Minh họa kiến trúc mô hình VidLoc [14]

MapNet+ và MapNet+PGO [9] mang cùng một kiến trúc mạng với MapNet trích xuất đặc trưng qua mạng ResNet34 và dùng một lớp tổng hợp trung bình toàn cục. Không chỉ dùng mắt mát vị trí tuyệt đối, mắt mát đo lường cảm biến trực quan cũng được tính toán để cải thiện hiệu suất dự đoán vị trí của MapNet. Phương pháp này cũng đồng thời tích hợp dữ liệu GNSS và IMU để giúp cải thiện hồi quy vị trí. Điều này giúp kết hợp dữ liệu đã được gắn nhãn và dữ liệu chưa gắn nhãn từ VO hay cảm biến để phục vụ cho việc học tự giám sát và đã thể hiện hiệu suất tốt dưới những điều kiện khó khăn, thử thách như thay đổi môi trường ngoài, thiếu sáng,... .

$$l = l_{labelleddata} + l_{unlabelleddata}$$

Với mắt mát từ dữ liệu chưa gắn nhãn có thể được tính toán thông qua việc kết hợp vị trí máy ảnh tương đối  $v_i, j$  và VO  $\hat{v}_i, j$  hay các cảm biến khác như IMU và GNSS.

MapNet+PGO đã có thể cải thiện hiệu suất đồng thời giảm thiểu chi phí tính toán thông qua việc sử dụng cải thiện đồ thị vị trí (PGO) để kết hợp kết quả vị trí MapNet+ và VO.

## Kết luận

Xét về các phương pháp mang hướng hồi quy vị trí tuyệt đối thông qua một ảnh duy nhất, các nghiên cứu có xu hướng tiến tới việc hàm mắt mát tự động hóa, không sử dụng siêu tham số và mang nhiều thông tin hơn để giảm việc sử dụng các tham số cố định. Khởi đầu từ PoseNet với việc sử dụng một hàm mắt mát cố định để tính tổng độ dời và hướng quay sử dụng một số tham số cân bằng. Sau đó, một hàm mắt mát có trọng số học được [46, 10] đã được đề xuất bằng cách thêm độ không đảm bảo phuong sai đồng nhất để tự động cân bằng mắt mát độ dời và hướng quay, tránh sử dụng siêu tham số đồng thời vượt qua hiệu suất của phương pháp sử dụng hàm mắt mát cố định. Ngoài phương pháp hàm mắt mát cố định và hàm mắt mát với trọng số học được, một số công trình [20, 11] đề xuất sử dụng mắt mát lỗi tái chiếu và mắt mát GPoseNet để thêm các định dạng thông tin khác như phân phối xác suất của vị trí - hướng quay đầu ra để cải thiện hàm mắt mát.

Với việc sử dụng nhiều ảnh cũng như học kết hợp các tác vụ phụ, các mô hình đã có thể không chỉ thu được kết quả định vị mà còn có được thông tin khác như VO, thông tin ngữ nghĩa, etc. MapNet, VlocNet, AtLocPlus [46] kết hợp cả hàm mắt mát vị trí tương đối và vị trí tuyệt đối để cải thiện hiệu quả tác vụ hồi quy. LSG [49] áp dụng một ràng buộc chuyển động vào hàm mắt mát trong khi đó VlocNet++ [30] thì đề xuất thêm ràng buộc ngữ nghĩa vào. DGRNet [23] kết hợp CTC và MSE vào quá trình tính toán hàm mắt mát.

Cho rằng các phương pháp hồi quy vị trí tuyệt đối thông qua ảnh đang bỏ phí giá trị của các ràng buộc thời gian, một số công trình [14, 9] đề xuất việc tận dụng các đoạn phim làm đầu vào cho tác vụ hồi quy vị trí. VidLoc [14] sử dụng RNN hai chiều để hồi quy vị trí 6DoF của máy ảnh. MapNet+ và MapNet+PGO [9] chủ yếu tận dụng VO vào hàm mắt mát để tăng cường hiệu quả tác vụ hồi quy.

### 2.3.2 Hồi quy vị trí tương đối - Relative Pose Regression

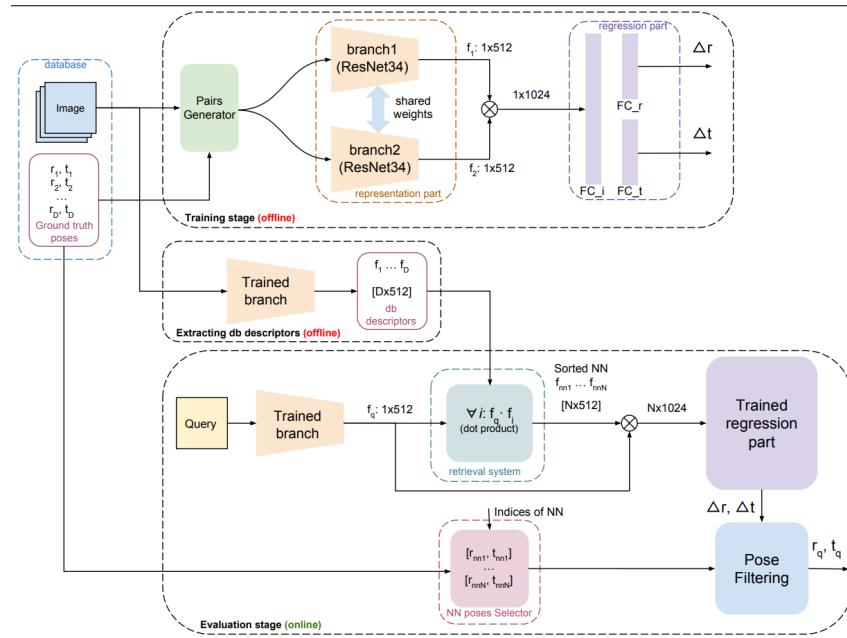
Mô hình hồi quy vị trí tuyệt đối học cách ánh xạ các pixel của ảnh sang vị trí của máy ảnh, thường được quyết định bởi hệ trục tọa độ của chính cảnh vật cụ thể mà máy ảnh đang chụp. Khác với hồi quy vị trí tuyệt đối, các phương pháp mang hướng tiếp cận hồi quy vị trí tương đối (Relative Pose Regression) chỉ tính toán vị trí tương đối của ảnh và thường được huấn luyện trên những tập dữ liệu đa cảnh để tăng cường khả năng mở rộng mô hình đầu cuối.

#### Hồi quy vị trí tương đối thông qua truy xuất rõ ràng - Relative camera pose regression through explicit retrieval

Quy trình hồi quy vị trí tương đối của máy ảnh có thể được hiểu như một quy trình bao gồm truy xuất ảnh có độ tương đồng cao nhất trong kho dữ liệu với ảnh nhận đầu vào và sau đó dự đoán vị trí tương đối giữa chúng để lấy được vị trí tuyệt đối của ảnh nhận đầu vào. Cho một ảnh  $I_a^c$  được chụp từ máy ảnh  $c$ , thông qua các phương pháp truy xuất ảnh từ kho dữ liệu, chúng ta có được ảnh có độ tương đồng cao nhất  $I_b^c$ . Nếu có được vị trí nền tảng  $p_b$  của ảnh  $I_b^c$  và vị trí tương đối giữa hai ảnh là  $p_{a \rightarrow b}$ , vị trí tuyệt đối  $p_a$  của ảnh đầu vào  $I_a^c$  có thể được xác định bằng các phép biến đổi toán học.

NNnet [22] là công trình đầu tiên đề xuất một phương pháp hồi quy vị trí tương đối dựa trên truy xuất ảnh. Đầu vào của phương pháp này là một ảnh và một kho dữ liệu ảnh có bao gồm vị trí nền tảng. Một tập các cặp ảnh được tận dụng để hồi quy vị trí tương đối thông qua một mạng Siamese với hai nhánh ResNet34 đã được hiệu chỉnh và một hàm mắt mát cố định. Ảnh có độ tương đồng gần nhất với ảnh nhận đầu vào có thể được tính toán xác định thông qua bộ trích xuất đặc trưng hình thành bởi

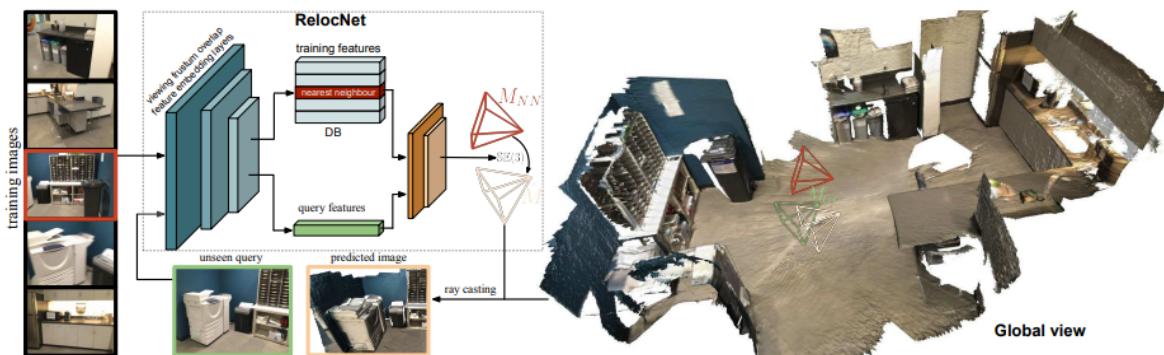
nhánh mạng CNN, sau đó vị trí tương đối và vị trí nền tảng của ảnh trích xuất sẽ kết hợp để tính toán xác định vị trí tuyệt đối của ảnh đầu vào.



Hình 2.17: Minh họa kiến trúc mô hình NNet [22]

RelocNet [5] cải tiến NNet với việc học liên tục thuộc đo với mục đích học các đặc trưng ảnh toàn cục với một góc nhìn chéo cự của máy ảnh để cải thiện kết quả, mất mát vị trí tương đối cũng được áp dụng. Mất mát vị trí tương đối học sự khác biệt vị trí giữa hai ma trận vị trí bằng cách sử dụng một biểu diễn ma trận cho hướng quay và độ dời vị trí. Hàm mất mát huấn luyện được định nghĩa như sau:

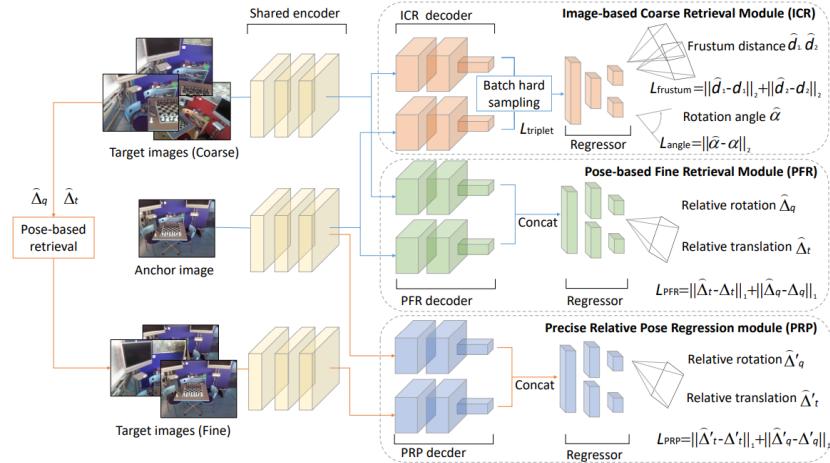
$$l = \alpha l_{SE(3)} + \beta l_{frustum}$$



Hình 2.18: Minh họa kiến trúc mô hình RelocNet [5]

Để giải quyết vấn đề hiệu suất giới hạn của các phương pháp hồi quy vị trí tương đối tiên nhiệm do sử dụng cùng đặc trưng cho cả hai bước truy xuất và hồi quy, CamNet [15] đề xuất một quy trình chia làm ba bước: truy xuất thô, truy xuất mịn và hồi quy vị trí. Kiến trúc mô hình được xây dựng dựa trên kiến trúc Siamese với ba nhánh mỗi bước. Kiến trúc thô-sang-mịn này đã mang lại những cải tiến về hiệu suất hồi quy cũng như khả năng mở rộng. Hàm mất mát của CamNet lấy ý tưởng dựa trên RelocNet, được định nghĩa như sau:

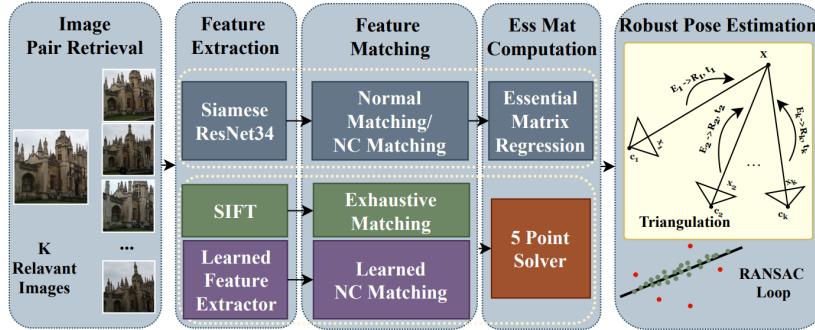
$$l = l_{frustum} + l_{angle} + l_{triplet} + l_{PFR} + l_{PRP}$$



Hình 2.19: Minh họa kiến trúc mô hình CamNet [15]

Qunjie Zhou và những cộng sự [50] sau khi phân tích các phương pháp hồi quy vị trí dựa trên việc truy xuất ảnh đã đề xuất một kiến trúc mới sử dụng ma trận thiết yếu và RANSAC để tính toán vị trí tuyệt đối. Một mạng Siamese ResNet34 với một lớp tìm sự tương ứng cố định (EssNet) và một lớp tìm sự tương ứng đồng thuận lân cận (NC-EssNet) được học để tạo ra một bản đồ điểm tương ứng phục vụ cho mục đích hồi quy về sau của ma trận thiết yếu. Hàm mất mát cải tiến khoảng cách Euclidean giữa ma trận thiết yếu với hai véc-tơ 9 chiều.

$$l_{ess}(E^*, E) = \|e - e^*\|_2$$

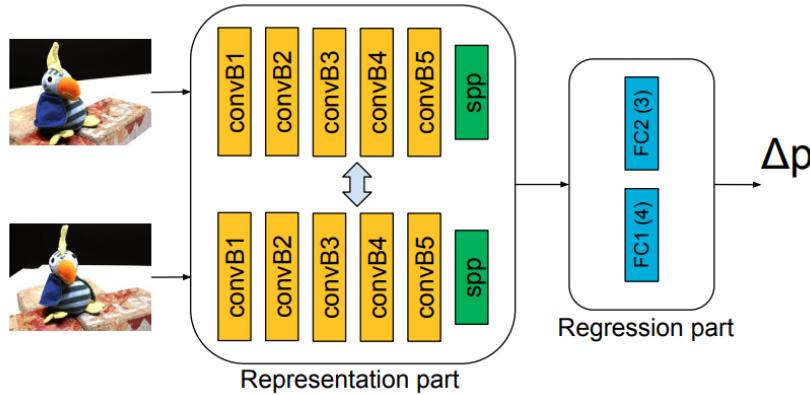


Hình 2.20: Minh họa kiến trúc mô hình EssNet [50]

### Hồi quy vị trí tương đối thông qua mạng CNN ngầm - Relative camera pose regression through implicit CNN

Để tránh việc phải thu thập và xây dựng một kho dữ liệu khổng lồ cũng như tốn kém thời gian thử nghiệm, một số phương pháp tìm cách hồi quy vị trí tương đối của máy ảnh thông qua một mạng Nơ-ron tích chập ngầm.

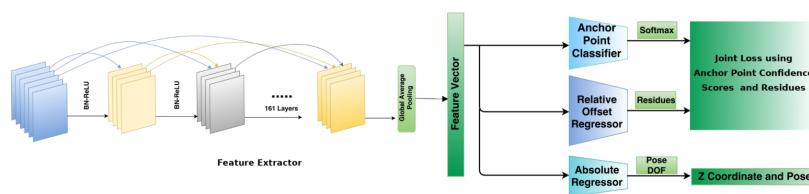
Relative NN [26] đề xuất một phương pháp đầu-cuối để hồi quy vị trí tương đối giữa hai máy ảnh bằng hai ảnh đầu vào. Kiến trúc mô hình là một mạng Nơ-ron hỗn hợp Siamese hai nhánh sử dụng mạng AlexNet đã được huấn luyện từ trước được dùng cho việc hồi quy với hàm mất mát Euclidean cố định.



Hình 2.21: Minh họa kiến trúc mô hình Relative Neural Network [26]

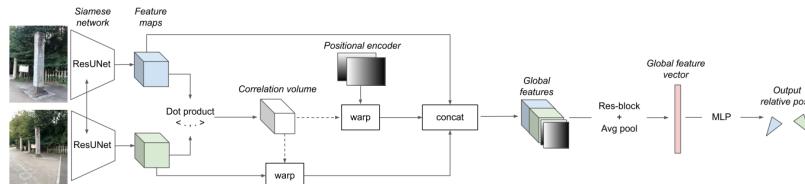
AnchorNet [31] tìm cách khắc phục vấn đề định vị bằng cách định nghĩa các địa danh thành các điểm mốc để học các điểm mốc tương đối của ảnh đầu vào cũng như độ lệch của chúng. Mô hình đa nhiệm bao gồm việc phân loại hình ảnh truy vấn đầu vào theo các điểm mốc cụ thể và tìm sự chênh lệch so với điểm mốc đã phân loại, điều này dẫn đến việc hình thành hàm mất mát.  $\hat{C}$ ,  $X$ , và  $Y$  đại diện cho kết quả phân loại và sự chênh lệch với sự thật nền tảng.

$$l = \sum_i [(X_i - \hat{X}_i)^2 + (Y_i - \hat{Y}_i)^2] \hat{C}^i$$



Hình 2.22: Minh họa kiến trúc mô hình AnchorNet [31]

Nhận thấy các phương pháp tái định vị trực quan (Visual Relocalization) hiện tại đều cần một kho dữ liệu ảnh khổng lồ nhằm mục đích xây dựng một mô hình 3 chiều cho một khung cảnh nhất định, một nhóm nghiên cứu [4] đề xuất phương pháp mang tên "Tái định vị không cần bản đồ (Map-free Relocalization)" với việc chỉ sử dụng duy nhất một ảnh làm đầu vào mà không cần phải xây dựng mô hình 3 chiều cho cảnh. [4] đã đề xuất hai phương pháp khác nhau để có thể xác định vị trí, góc quay chính xác từ một ảnh: thứ nhất là tận dụng ma trận thiết yếu kết hợp với các kỹ thuật tìm sự tương ứng giữa đặc trưng ảnh, thứ hai là hồi quy vị trí tương đối.



Hình 2.23: Minh họa kiến trúc mô hình hồi quy vị trí tương đối của Map-free [4]

## Kết luận

Để thực hiện tác vụ hồi quy vị trí tương đối, các phương pháp dựa trên truy xuất ảnh [22, 5, 15, 50] tận dụng một quy trình nhiều bước để lấy được vị trí tuyệt đối với bước truy xuất ảnh làm trọng tâm. NNet [22] là công trình đầu tiên đề xuất một phương pháp hồi quy vị trí tương đối dựa trên truy xuất ảnh, với RelocNet [5] tìm cách cải tiến NNet với việc học liên tục thước đo với mục đích học các đặc trưng ảnh toàn cục kết hợp góc nhìn chéo cụt của máy ảnh để cải thiện kết quả. CamNet [15] đề xuất một quy trình chia làm ba bước: truy xuất thô, truy xuất mịn và hồi quy vị trí với ba nhánh CNN mỗi bước và đã mang lại những cải tiến về hiệu suất hồi quy cũng như khả năng mở rộng.

Trong khi đó các phương pháp dựa trên CNN [26, 31, 4] lại đề xuất một hướng giải quyết để hồi quy trực tiếp vị trí tương đối ngay trong mạng nơ-ron. Relative NN [26] đề xuất một mô hình đầu-cuối Siamese hai nhánh để hồi quy vị trí tương đối giữa hai ảnh đầu vào. AnchorNet [31] tìm cách khắc phục vấn đề định vị bằng cách định nghĩa các địa danh thành các điểm mốc để học các điểm mốc tương đối của ảnh đầu vào cũng như độ lệch của chúng. Map-free [4] đề xuất việc không sử dụng ảnh để tái kiến trúc lại cảnh, đồng thời cung cấp hai phương hướng để hồi quy vị trí từ một ảnh.

### 2.3.3 Tái tạo kiến trúc từ chuyển động - Structure From Motion

## 2.4 Phân tích và tổng hợp

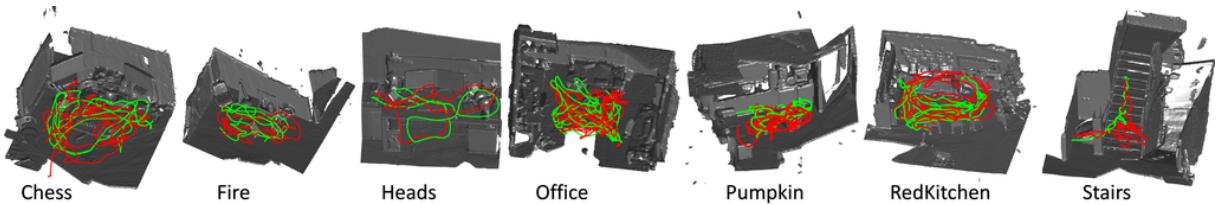
Về các phương pháp hồi quy vị trí máy ảnh, các nghiên cứu cho thấy hiệu suất dự đoán vị trí vẫn chưa thể vượt qua các phương pháp tái kiến trúc từ chuyển động. Các phương pháp hồi quy vị trí máy ảnh vẫn còn đang trên đà phát triển tuy nhiên vẫn gặp nhiều khó khăn với việc giải quyết các đặc trưng cục bộ trùng lặp cũng như việc tiêu tốn nhiều tài nguyên tính toán. Hầu hết công trình tập trung vào các điểm đặc trưng có độ chắc chắn cao hoặc mô tả đặc trưng chính xác dưới điều kiện mang tính thử thách cao. Ngoài ra, phần lớn các phương pháp được đề xuất chỉ có thể áp dụng trong một phạm vi nhỏ và vẫn chưa chứng tỏ được hiệu suất trên những tập dữ liệu có quy mô lớn.

## 2.5 Một số tập dữ liệu phổ biến được sử dụng

### 2.5.1 Tập dữ liệu trong không gian nhỏ

#### 7Scenes

Tập dữ liệu 7-Scenes [40] bao gồm các ảnh RGB-D thuộc bảy khung cảnh khác nhau được chụp từ một máy ảnh cầm tay Kinect RGB-D ở độ phân giải 640x480. Bảy khung cảnh bao gồm: "Chess", "Fire", "Heads", "Office", "Pumpkin", "RedKitchen" và "Stairs". Với mỗi cảnh sẽ có vài chuỗi khung ảnh RGB-D. Mỗi chuỗi bao gồm khoảng từ 1000 đến 5000 khung ảnh. Mỗi khung sẽ gồm: ảnh màu, độ sâu và vị trí.



Hình 2.24: Minh họa tập dữ liệu 7-Scenes [40]

### Cambridge Landmark

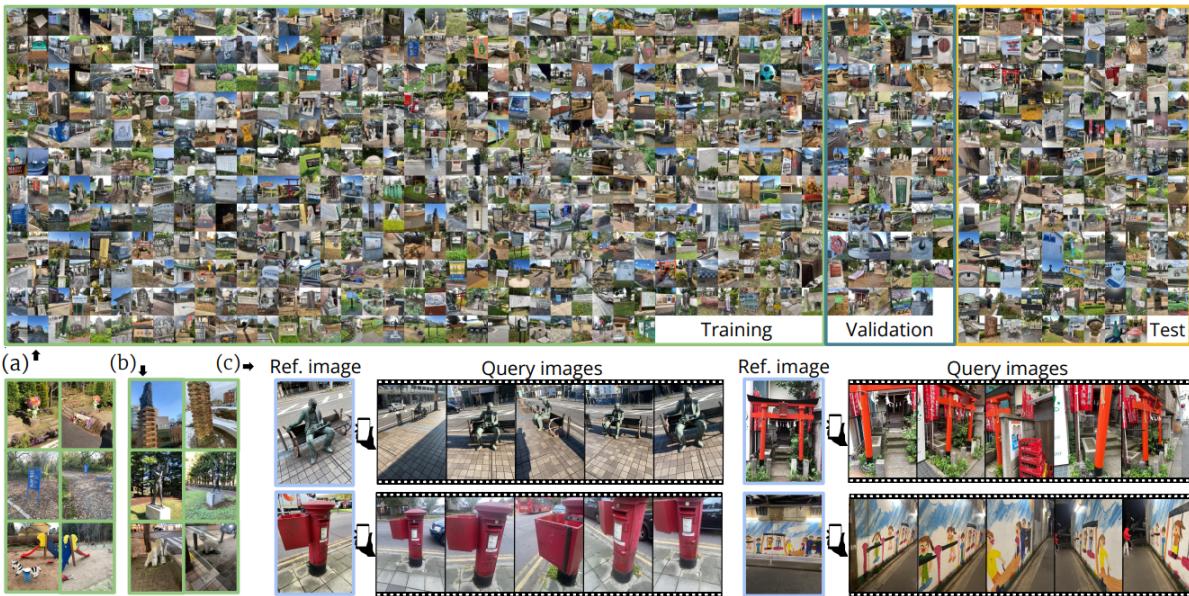
Tập dữ liệu Cambridge Landmarks [21] là một tập dữ liệu định vị thành thị bao gồm năm khung cảnh khác nhau. Các yếu tố dày đặc quan trọng như phương tiện giao thông hay người đi bộ cũng xuất hiện trong tập dữ liệu này, ngoài ra dữ liệu cũng được thu thập ở nhiều thời điểm trong ngày đại diện cho các yếu tố ánh sáng và điều kiện thời tiết khác nhau. Cambridge Landmarks được tạo ra nhờ vào việc áp dụng các kỹ thuật tái tạo kiến trúc từ chuyển động. Một chiếc điện thoại thông minh Google LG Nexus 5 được một người đi bộ trên phố sử dụng để ghi lại đoạn phim chất lượng cao cho mỗi cảnh. Mỗi đoạn phim sau đó sẽ được lấy mẫu với tần số 2Hz để trích xuất ảnh cho quy trình tái tạo kiến trúc từ chuyển động. Mỗi vị trí máy ảnh sẽ cách nhau khoảng 1m.



Hình 2.25: Minh họa tập dữ liệu Cambridge Landmarks [21]

### Niantic Map-free Relocalization Dataset

Tập dữ liệu Niantic Map-free Relocalization [4] là một tập dữ liệu được thu thập chủ yếu để giúp ích cho phương pháp định vị Map-free [4]. Tập dữ liệu bao gồm 655 cảnh bên ngoài với mỗi cảnh sẽ chứa một "địa điểm đáng chú ý" như một pho tượng, cổng, bảng hiệu,... sao cho địa điểm đó phải được xác định rõ trong một bức ảnh. Các cảnh được chia ra thành 460 cảnh phục vụ cho tác vụ huấn luyện, 65 cảnh phục vụ cho tác vụ kiểm tra quy trình huấn luyện và 130 cảnh phục vụ cho quá trình kiểm thử. Mỗi ảnh trong tập huấn luyện đều được gắn kèm vị trí tuyệt đối. Với tập kiểm thử và kiểm tra quy trình, mỗi cảnh sẽ được kèm theo một ảnh đại diện cũng như vị trí tuyệt đối tại cảnh. Ngoài ra, ma trận tham số nội tại của máy ảnh cũng được gắn kèm theo mỗi ảnh trong tập dữ liệu.



Hình 2.26: Minh họa tập dữ liệu Niantic Map-free Relocalization [4]

## 2.5.2 Tập dữ liệu thành thị

### Aachen Day-Night

Tập dữ liệu Aachen Day-Night [37] bao gồm 14.607 ảnh được chụp với nhiều máy ảnh khác nhau bao phủ cả thành phố Aachen thuộc quốc gia Đức. Các ảnh dữ liệu được chụp ở nhiều thời điểm trong ngày và trong năm, cụ thể là khoảng thời gian trong 2 năm. Hệ quả mang lại là tập dữ liệu bao phủ nhiều điều kiện ngoại cảnh như thời tiết, ánh sáng cũng như sự thay đổi của công trình kiến trúc trong khu vực.



Hình 2.27: Minh họa tập dữ liệu Aachen Day-Night [37]

### Pittsburgh 250k

Tập dữ liệu Pittsburgh 250k [43] là một tập dữ liệu tương đối rộng bao phủ thành phố Pittsburgh của Mỹ. Đây là một tập dữ liệu tương đối phổ biến trong thị giác máy tính, cụ thể là ở tác vụ nhận diện địa điểm trực quan, truy xuất ảnh và định vị trực quan.

## GSV-Cities

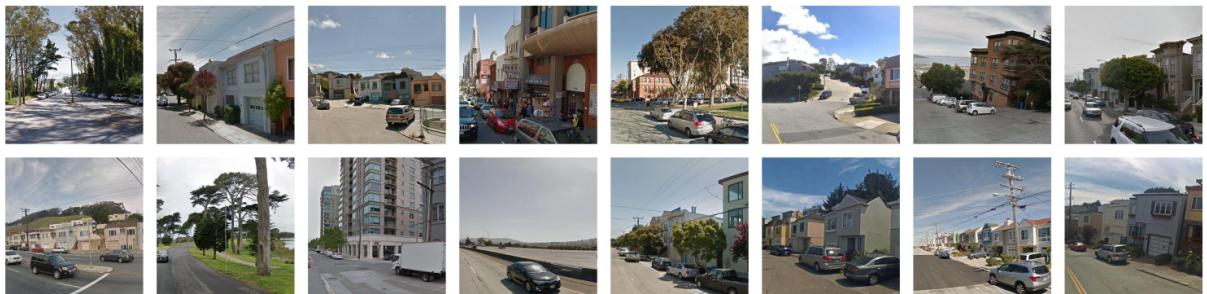
Tập dữ liệu GSV-Cities [1] bao phủ một vùng địa lý cực kỳ rộng lớn với hơn 40 thành phố xuyên lục địa trong khoảng thời gian 14 năm liên tục. GSV-Cities chứa khoảng hơn 530.000 ảnh - khoảng hơn 62.000 vị trí. Mỗi vị trí sẽ có khoảng từ 4 đến 20 ảnh. Đồng thời mỗi vị trí sẽ cách nhau một khoảng ít nhất 100m.



Hình 2.28: Minh họa tập dữ liệu GSV-Cities [1]

## SF-XL

Tập dữ liệu San Francisco Extra Large (SF-XL) [7] được tạo nên từ 3.43 triệu ảnh 360 độ thu thập từ kho ảnh Google Streetview. Các ảnh này sau đó được cắt ra thành 41.2 triệu ảnh. Mỗi ảnh cắt ra đều được gắn nhãn 6DoF (bao gồm cả GPS). Dữ liệu được thu thập từ năm 2009 đến năm 2021, dẫn đến việc tập dữ liệu bao phủ nhiều điều kiện ngoại cảnh như thời tiết, ánh sáng cũng như sự thay đổi của công trình kiến trúc trong khu vực.



Hình 2.29: Minh họa tập dữ liệu San Francisco Extra Large [7]

# Chương 3

## PHƯƠNG PHÁP ĐỀ XUẤT

### 3.1 Tổng quan về mô hình được đề xuất

Như một phương pháp hồi quy vị trí tương đối bình thường, mô hình được đề xuất sẽ bao gồm hai bộ phận chính là bộ phận truy xuất ảnh và bộ phận hồi quy vị trí tương đối dựa trên cặp ảnh. Cụ thể hơn

- Bộ phận truy xuất ảnh sẽ sử dụng mô hình được đề xuất trong bài báo nghiên cứu MixVPR của ...
- Bộ phận hồi quy vị trí tương đối dựa trên cặp ảnh sẽ sử dụng mô hình tương quan 2D-2D được đề xuất trong bài báo nghiên cứu Map-free Relocalization.

#### 3.1.1 MixVPR

**Ý tưởng đằng sau mô hình**

**Những bước xử lý của mô hình**

**Lý do chọn mô hình**

#### 3.1.2 Mô hình tương quan 2D-2D của Map-free Relocalization

**Ý tưởng đằng sau mô hình**

**Những bước xử lý của mô hình**

**Lý do chọn mô hình**

### 3.2 Tiêu chí đánh giá

#### 3.2.1 MixVPR

#### 3.2.2 Mô hình tương quan 2D-2D của Map-free Relocalization

# **Chương 4**

## **ĐO ĐẠC VÀ ĐÁNH GIÁ**

### **4.1 Mô hình MixVPR**

**Mô tả quá trình thực nghiệm**

**Kết quả thực nghiệm**

**Nhận xét**

### **4.2 Mô hình Map-free Relocalization**

**Mô tả quá trình thực nghiệm**

**Kết quả thực nghiệm**

**Nhận xét**

### **4.3 Hướng phát triển**

### **4.4 Kết luận**

# **Chương 5**

## **KẾ HOẠCH TƯƠNG LAI**

**5.1 Thành quả đạt được**

**5.2 Kế hoạch luận văn tốt nghiệp**

# Tài liệu tham khảo

- [1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, November 2022.
- [2] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition, 2023.
- [3] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2016.
- [4] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image, 2022.
- [5] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 782–799, Cham, 2018. Springer International Publishing.
- [6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part I* 9, pages 404–417. Springer, 2006.
- [7] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geolocation for large-scale applications, 2022.
- [8] Eric Brachmann and Carsten Rother. Visual camera re-localization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021.
- [9] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization, 2018.
- [10] Mai Bui, Christoph Baur, Nassir Navab, Slobodan Ilic, and Shadi Albarqouni. Adversarial networks for camera pose regression and refinement, 2019.
- [11] Mingpeng Cai, Chunhua Shen, and Ian D. Reid. A hybrid probabilistic model for camera relocalization. In *British Machine Vision Conference*, 2018.
- [12] Mohamed Chaabane, Lionel Gueguen, Ameni Trabelsi, Ross Beveridge, and Stephen O’Hara. End-to-end learning improves static object geo-localization from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2063–2072, 2021.
- [13] Boris Chidlovskii and Assem Sadek. Adversarial transfer of pose estimation regression, 2020.
- [14] Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, and Hongkai Wen. Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization, 2017.

- [15] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera re-localization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2871–2880, 2019.
- [16] Martin Humenberger, Gabriela Csurka Khedari, Nicolas Guerin, and Boris Chidlovskii. Methods for visual localization. <https://europe.naverlabs.com/blog/methods-for-visual-localization/>. Accessed: 2023-12-05.
- [17] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2136–2145, 2017.
- [18] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *arXiv preprint arXiv:2308.00688*, 2023.
- [19] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization, 2016.
- [20] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning, 2017.
- [21] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization, 2016.
- [22] Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. Camera relocalization by computing pairwise relative poses using convolutional neural network, 2017.
- [23] Yimin Lin, Zhaoxiang Liu, Jianfeng Huang, Chaopeng Wang, Guoguang Du, Jinqiang Bai, Shiguo Lian, and Bill Huang. Deep global-relative networks for end-to-end 6-dof visual localization and odometry, 2019.
- [24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.
- [25] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Image-based localization using hourglass networks, 2017.
- [26] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks, 2017.
- [27] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. Scalable 6-dof localization on mobile devices. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part II 13*, pages 268–283. Springer, 2014.
- [28] Tayyab Naseer and Wolfram Burgard. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. pages 1525–1530, 09 2017.
- [29] Guohao Peng, Yufeng Yue, Jun Zhang, Zhenyu Wu, Xiaoyu Tang, and Danwei Wang. Semantic reinforced attention learning for visual place recognition. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13415–13422. IEEE, 2021.
- [30] Noha Radwan, Abhinav Valada, and Wolfram Burgard. Vlocnet++: Deep multi-task learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 3(4):4407–4414, October 2018.
- [31] Soham Saha, Girish Varma, and C. V. Jawahar. Improved visual relocalization by discovering anchor points, 2018.
- [32] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceed-*

- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019.
- [33] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kortschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21632–21642, 2023.
  - [34] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016.
  - [35] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018.
  - [36] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1646, 2017.
  - [37] Torsten Sattler, Tobias Weyand, B. Leibe, and Leif P. Kobbelt. Image retrieval for image-based localization revisited. In *British Machine Vision Conference*, 2012.
  - [38] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers, 2021.
  - [39] Yoli Shavit, Ron Ferens, and Yosi Keller. Coarse-to-fine multi-scene pose regression with transformers. *IEEE transactions on pattern analysis and machine intelligence*, PP, 08 2023.
  - [40] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013.
  - [41] Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Robotics: Science and Systems XI*, pages 1–10, 2015.
  - [42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.
  - [43] Akihiko Torii, Josef Sivic, Tomá Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 883–890, 2013.
  - [44] Abhinav Valada, Noha Radwan, and Wolfram Burgard. Deep auxiliary learning for visual localization and odometry, 2018.
  - [45] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation, 2017.
  - [46] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization, 2019.

- 
- [47] Jian Wu, Liwei Ma, and Xiaolin Hu. Delving deeper into convolutional neural networks for camera relocalization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5644–5651, 2017.
  - [48] Meng Xu, Youchen Wang, Bin Xu, Jun Zhang, Jian Ren, Stefan Poslad, and Pengfei Xu. A Critical Analysis of Image-based Camera Pose Estimation Techniques. *arXiv e-prints*, page arXiv:2201.05816, January 2022.
  - [49] Fei Xue, Xin Wang, Zike Yan, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Local supports global: Deep camera relocalization with sequence enhancement, 2019.
  - [50] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices, 2020.