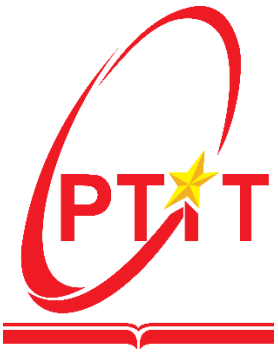
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



BÁO CÁO HÀNG TUẦN

HỌC PHẦN: THỰC TẬP CƠ SỞ

ĐỀ TÀI:

NGHIÊN CỨU ƯỚC LƯỢNG KHOẢNG CÁCH BẰNG CAMERA 2D

Giảng viên hướng dẫn: TS. Kim Ngọc Bách

Sinh viên thực hiện:

B22DCCN634 Trần Hữu Phúc

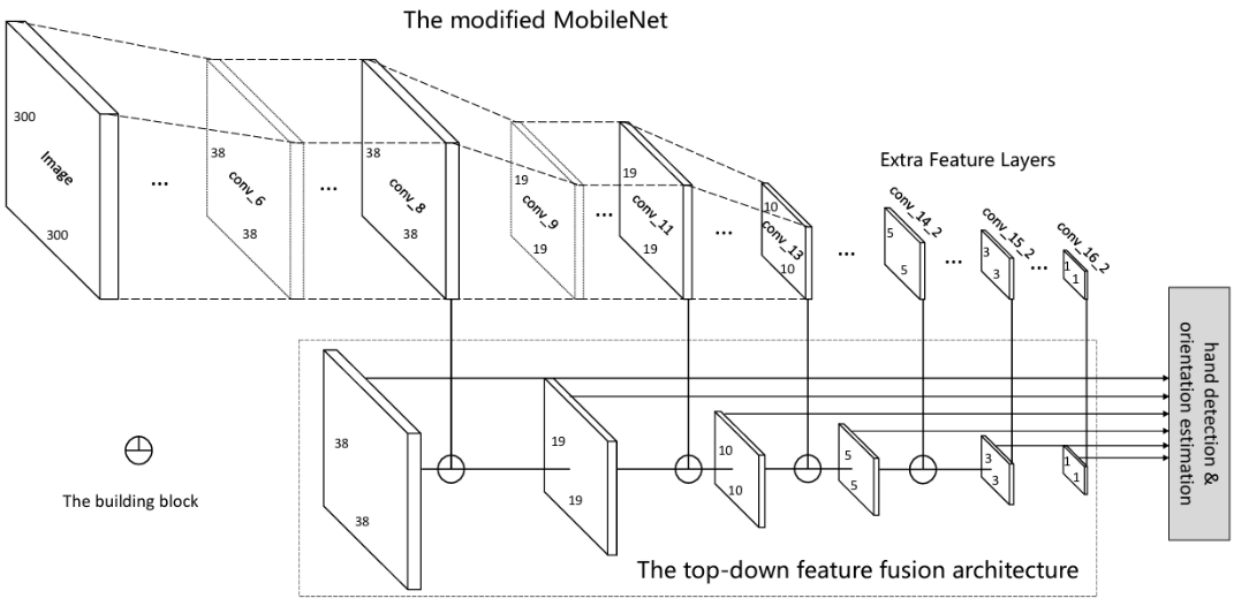
15-22/03/2025

A. BÁO CÁO TIẾN ĐỘ

1.1. Các nghiên cứu nước ngoài về phát hiện bàn tay

1.1.1. Phương pháp phát hiện và ước lượng tư thế tay dựa trên CNN

Em đã nghiên cứu bài báo "A Light CNN Based Method for Hand Detection and Orientation Estimation" của các tác giả Li Yang và các cộng sự. Bài báo giới thiệu một phương pháp dựa trên mạng nơ-ron tích chập (CNN) để phát hiện và ước tính hướng bàn tay trong ảnh RGB, tập trung vào việc cân bằng giữa độ chính xác và hiệu suất xử lý. Phương pháp sử dụng cấu trúc Single Shot Multibox Detector (SSD) kết hợp với MobileNet đã được tùy chỉnh, tạo ra sáu bản đồ đặc trưng với các độ phân giải khác nhau, hỗ trợ phát hiện bàn tay ở nhiều kích thước. Để cải thiện khả năng nhận diện trong các trường hợp khó khăn như bàn tay nhỏ hoặc bị che khuất, một kiến trúc hợp nhất đặc trưng từ trên xuống được áp dụng, tích hợp thông tin ngữ cảnh trên nhiều mức độ.



Kiến trúc mạng tổng thể của phương pháp.

Hình trên mô tả mô hình mạng nơ-ron sâu dựa trên MobileNet dùng để phát hiện bàn tay và ước lượng hướng bàn tay.

Cấu trúc chính:

- Modified MobileNet: Trích xuất đặc trưng từ hình ảnh đầu vào.
- Hợp nhất đặc trưng: Kết hợp thông tin từ nhiều lớp để cải thiện hiệu quả.
- Lớp bổ sung: Tăng độ chính xác với các đặc trưng sâu hơn.
- Đầu ra: Dự đoán vị trí và hướng bàn tay.

Việc phát hiện bàn tay được thực hiện thông qua dự đoán xác suất và vị trí các hộp giới hạn mặc định, sau đó lọc kết quả bằng thuật toán Non-Maximum Suppression. Phương pháp cũng ước tính hướng bàn tay bằng cách xác định hộp giới hạn xoay qua hai vector vuông góc đại diện cho trục chính và phụ. Thay vì dự đoán trực tiếp góc hoặc vector, mô hình tính toán các phép chiếu của vector lên trục ngang và dọc, kết hợp với vị trí cổ tay để xác định hướng.

Huấn luyện mô hình trên bộ dữ liệu Oxford Hand Dataset, phương pháp sử dụng các kỹ thuật tăng cường dữ liệu và tối ưu hóa hàm mất mát bao gồm các thành phần phân loại, định vị, và ước tính vector. Kết quả cho thấy mô hình đạt độ chính xác trung bình (Average Precision - AP) 83.2%, vượt trội hơn các phương pháp trước đó (Le et al.: 75.1%, Deng et al.: 57.7%) và tốc độ xử lý 139 fps trên GPU Nvidia Titan X, nhanh hơn gần 30 lần so với phương pháp tốt nhất trước đó. Mô hình cũng đạt 63.36% độ chính xác trong việc ước tính hướng bàn tay với sai lệch không quá 10°.

So sánh hiệu suất phát hiện bàn tay bằng nhiều phương pháp khác nhau.

Phương pháp	AP	Thời gian	Môi trường thử nghiệm
Mittal et al.	48.2%	2 phút	quad-core 2.5 GHz CPU
Deng et al.	57.7%	0.1 giây	quad-core 2.9 GHz CPU with

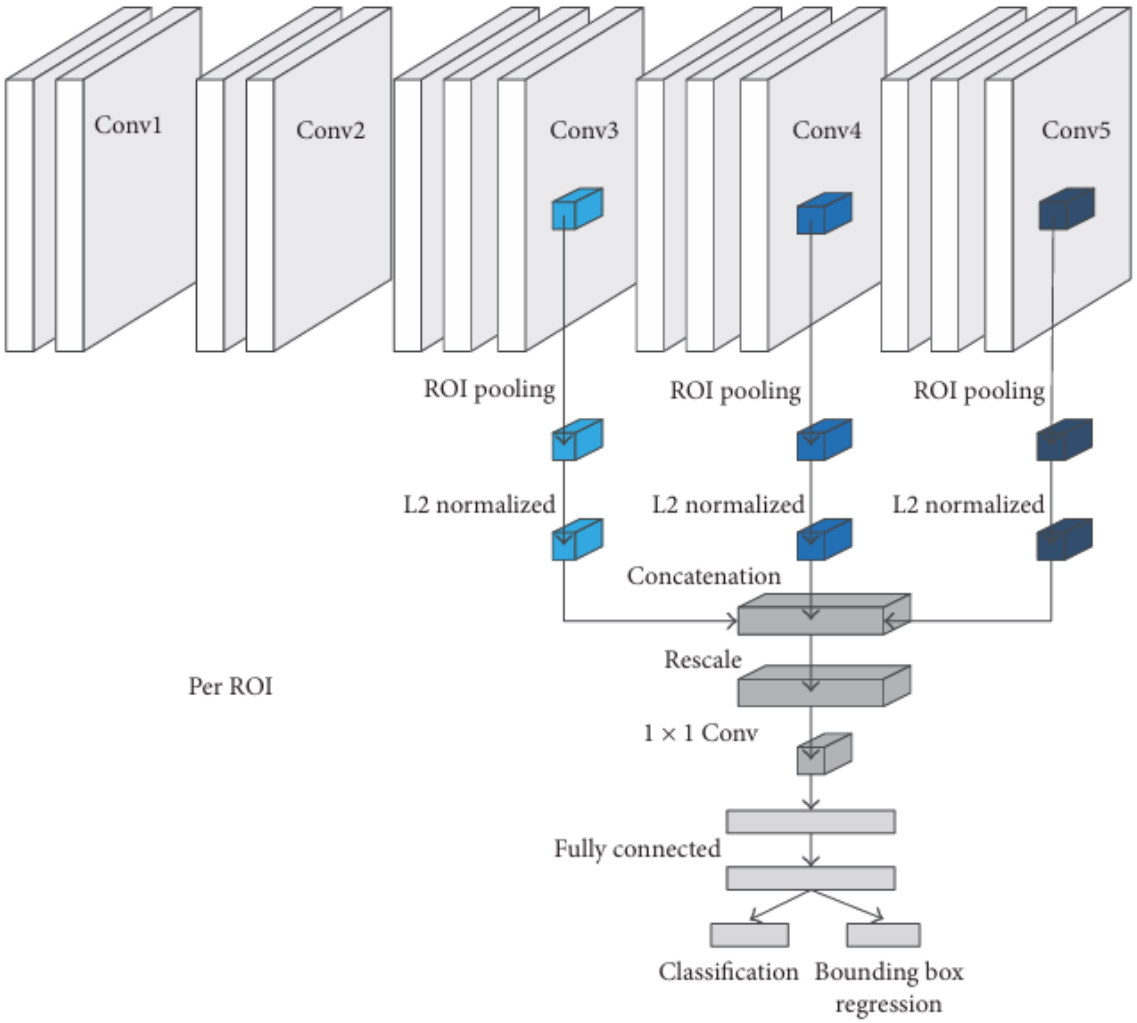
			Nvidia Titan X
Le et al.	75.1%	0.2150 giây	6 cores 3.5 GHz CPU with Nvidia Titan X
This method	83.2%	0.0072 giây	8 cores 3.0 GHz CPU with Nvidia Titan X

Bảng trên so sánh hiệu suất phát hiện bàn tay giữa phương pháp được đề xuất và các phương pháp trước đó. Kết quả cho thấy phương pháp của bài báo đạt độ chính xác trung bình (AP) cao nhất với 83.2%, vượt xa các phương pháp Mittal et al. (48.2%), Deng et al. (57.7%), và Le et al. (75.1%). Về thời gian xử lý, phương pháp này cũng đạt hiệu suất vượt trội với chỉ 0.0072 giây, nhanh hơn đáng kể so với Le et al. (0.2150 giây) và Deng et al. (0.1 giây), đồng thời rút ngắn thời gian so với Mittal et al. vốn yêu cầu tới 2 phút trên CPU.

Phương pháp này đánh dấu một bước tiến quan trọng trong việc phát hiện và ước tính hướng bàn tay, với tiềm năng ứng dụng lớn trong các lĩnh vực như nhận diện cử chỉ và tương tác giữa người và máy.

1.1.2. Mạng nơ-ron tích chập đa tỉ lệ cho phát hiện bàn tay

Shiyang Yan và các cộng sự đề xuất một phương pháp phát hiện bàn tay trong ảnh tĩnh, sử dụng mô hình học sâu đa tỉ lệ để xử lý các vấn đề về biến đổi kích thước và ngữ cảnh phức tạp. Phương pháp này dựa trên mô hình VGG16 được cải tiến, kết hợp các đặc trưng từ nhiều lớp tích chập để tận dụng thông tin đa tỉ lệ. Hệ thống sử dụng cơ chế Region of Interest (RoI) pooling từ các lớp tích chập khác nhau, cho phép phát hiện các bàn tay nhỏ và xử lý hiệu quả các đối tượng có sự biến đổi lớn về hình dạng và kích thước.



Cấu trúc mô hình của các mạng.

Hình trên mô tả một kiến trúc mạng nơ-ron sâu sử dụng cơ chế RoI pooling để phát hiện đối tượng. Đầu vào được trích xuất đặc trưng qua các lớp Conv1 đến Conv5. Các vùng quan tâm (RoIs) được lấy từ các lớp Conv3, Conv4 và Conv5 thông qua RoI pooling, sau đó chuẩn hóa bằng L2 normalization. Đặc trưng từ các RoIs được kết hợp, chuẩn hóa kích thước (Rescale), và xử lý qua một lớp convolution 1x1 trước khi đưa vào các lớp fully connected. Đầu ra cuối cùng gồm phân loại nhãn (Classification) và dự đoán vị trí hộp giới hạn (Bounding box regression). Kiến trúc này tận dụng thông tin từ nhiều cấp độ đặc trưng để tăng hiệu quả phát hiện.

Cụ thể, thay vì chỉ sử dụng lớp cuối cùng của CNN để biểu diễn đặc trưng, phương pháp này tích hợp các đặc trưng từ các lớp Conv3, Conv4, và Conv5 của VGG16. Kỹ thuật chuẩn hóa L2 được áp dụng để cân bằng giá trị đặc trưng giữa các lớp, kết hợp với các lớp tích chập 1x1 để giảm chiều dữ liệu. Phương pháp sau đó thực hiện phân loại và hồi quy vị trí hộp giới hạn thông qua các lớp fully connected, sử dụng framework Fast R-CNN làm nền tảng.

Hệ thống được đánh giá trên hai bộ dữ liệu benchmark là Oxford Hand Detection Dataset và VIVA Hand Detection Challenge. Trên Oxford dataset, mô hình đạt độ chính xác trung bình (AP) 58.4%, vượt qua các phương pháp trước đó (Mittal et al.: 48.2%, VGG16 baseline: 56.8%). Trên VIVA dataset, phương pháp đạt AP 92.8% ở cấp độ dễ (L1) và 84.7% ở cấp độ khó (L2), trong đó mức L2 bao gồm cả các bàn tay nhỏ và góc nhìn phức tạp. Các kết quả này khẳng định hiệu quả vượt trội của việc sử dụng thông tin đa tỉ lệ trong phát hiện bàn tay, đặc biệt đối với các đối tượng có kích thước nhỏ.

Độ chính xác trung bình (AP) trên bộ dữ liệu VIVA L1 và L2 và so sánh với các phương pháp trước đó.

Phương pháp	L1 set	L2 set
CNNRegionSampling	66.8%	57.8%
ACFDepth4	70.1%	60.1%
YOLO	76.4%	69.5%
FRCNN	90.7%	**86.5%**
Mô hình đề xuất (Multiscale Fast R-CNN)	**92.8%**	8 4. 7%

Độ chính xác trung bình (AP) trên bộ dữ liệu phát hiện bàn tay Oxford và so sánh với các phương pháp trước đó. (Chỉ các trường hợp bàn tay lớn).

Phương pháp	AP
Multiple proposals	48.2%
VGG16	56.8%
Mô hình đề xuất	58.4%

Bảng 2 và Bảng 3 cung cấp so sánh hiệu suất của mô hình đa tỉ lệ được đề xuất với các phương pháp trước đó trên hai bộ dữ liệu khác nhau. Trên bộ dữ liệu VIVA, mô hình đạt độ chính xác trung bình (AP) cao nhất ở cấp độ dễ (L1) với 92.8%, vượt trội so với YOLO (76.4%) và Fast R-CNN (90.7%). Ở cấp độ khó (L2), mô hình đạt AP 84.7%, chỉ thấp hơn một chút so với Fast R-CNN (86.5%) nhưng vẫn vượt xa các phương pháp khác như YOLO (69.5%). Trên bộ dữ liệu phát hiện bàn tay Oxford, mô hình đạt AP 58.4%, cao hơn so với VGG16 (56.8%) và phương pháp Multiple Proposals (48.2%). Những kết quả này chứng minh ưu điểm của việc sử dụng thông tin đa tỉ lệ, đặc biệt trong việc phát hiện các bàn tay nhỏ hoặc trong các tình huống ngữ cảnh phức tạp.

Tuy nhiên, phương pháp này vẫn gặp khó khăn trong việc phân biệt các đối tượng có hình dáng tương tự bàn tay như cánh tay, bàn chân, hoặc các chi tiết trên quần áo. Các tác giả đề xuất rằng việc tích hợp thêm thông tin ngữ cảnh có thể giúp cải thiện khả năng phân biệt này trong tương lai.

1.2. Nghiên cứu trong nước về phát hiện bàn tay

Em đã nghiên cứu bài báo “Nhận dạng cử chỉ bàn tay dùng mạng nơ-ron chập” của nhóm nghiên cứu trường Đại học Sư phạm thành phố Hồ Chí Minh.

Trong bài báo này, nhóm tác giả trình bày một thiết kế mạng nơ-ron tích chập cho bài toán nhận dạng cử chỉ bàn tay. Dữ liệu hình ảnh cử chỉ tay được thu thập qua camera và trải qua các bước tiền xử lý, bao gồm chuyển đổi sang thang xám để giảm độ phức tạp, loại bỏ nhiễu và chuẩn hóa kích thước ảnh nhằm đảm bảo tính nhất quán trong đầu vào cho mô hình.

Hệ thống CNN được thiết kế với các lớp chính như lớp tích chập (convolutional layers) để trích xuất đặc trưng không gian, tập trung vào các yếu tố như đường viền và hình dạng của cử chỉ tay. Sau đó, lớp gộp (pooling layers) được sử dụng để giảm kích thước dữ liệu, giúp giảm thiểu số lượng tham số và cải thiện hiệu quả tính toán. Cuối cùng, các lớp kết nối hoàn toàn (fully connected layers) chịu trách nhiệm phân loại các cử chỉ thành các nhãn tương ứng. Thuật toán tối ưu hóa Stochastic Gradient Descent (SGD) và hàm mất mát được sử dụng trong quá trình huấn luyện để cập nhật trọng số, cải thiện độ chính xác của dự đoán. Dữ liệu được chia thành tập huấn luyện và kiểm tra để đảm bảo đánh giá mô hình một cách khách quan.

Kết quả cho thấy hệ thống đạt được độ chính xác trung bình rất cao, lên đến 98.6% trong điều kiện thích hợp, vượt trội so với các phương pháp truyền thống. Đặc biệt, trong môi trường nhiều nhiễu hoặc điều kiện ánh sáng không ổn định, mô hình vẫn duy trì được hiệu suất ổn định, chứng minh khả năng thích ứng tốt với các biến động thực tế. Các chỉ số như độ nhạy (sensitivity) và độ đặc hiệu (specificity) cũng được phân tích chi tiết, khẳng định rằng hệ thống không chỉ phân loại chính xác các cử chỉ phổ biến mà còn nhận diện tốt các trường hợp khó.

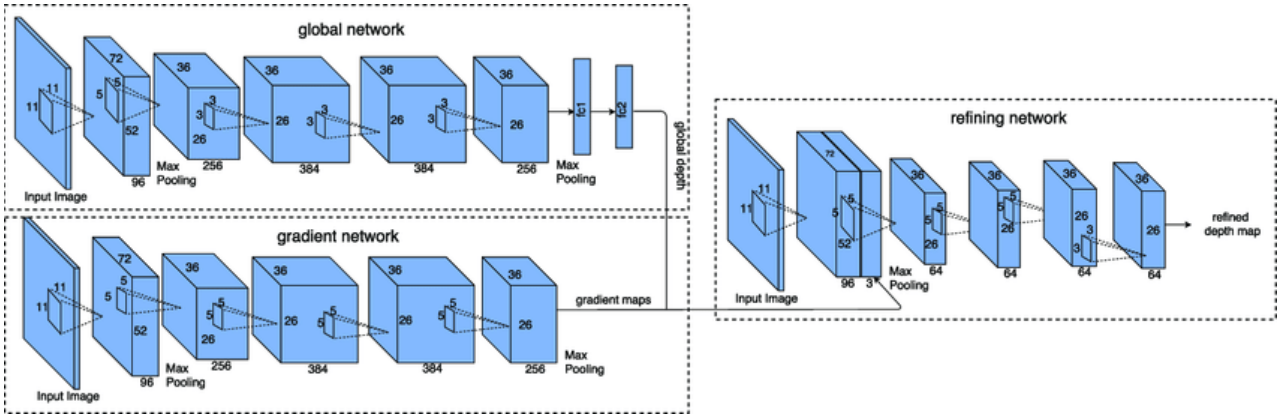
Ngoài ra, so với các phương pháp trước đây, hệ thống này mang lại lợi thế nhờ khả năng khai thác đặc trưng không gian hiệu quả từ hình ảnh, đồng thời có thể mở rộng ứng dụng trong các lĩnh vực như điều khiển thiết bị thông qua cử chỉ, giao tiếp không chạm, hoặc hỗ trợ các giải pháp thông minh trong nhà. Kết quả

nghiên cứu nhấn mạnh tiềm năng của CNN trong việc cải thiện hiệu quả các hệ thống nhận diện cử chỉ tay hiện nay.

1.3. Các phương pháp ước lượng khoảng cách đối với ảnh RGB

1.3.1. Mạng ước lượng độ sâu

Mạng ước lượng độ sâu (Depth Estimation Networks) là phương pháp sử dụng học sâu (deep learning) để dự đoán bản đồ chiều sâu (depth map) từ ảnh RGB. Đây là cách tiên tiến để suy luận thông tin về khoảng cách từ ảnh 2D, vốn không chứa dữ liệu độ sâu trực tiếp.



Cấu trúc mạng ước lượng độ sâu.

Kiến trúc mạng nơ-ron sâu gồm ba thành phần chính: mạng toàn cục (global network), mạng gradient (gradient network) và mạng tinh chỉnh (refining network). Mạng toàn cục trích xuất các đặc trưng tổng thể từ ảnh đầu vào, tạo ra bản đồ độ sâu thô. Mạng gradient tập trung vào việc phân tích các đặc trưng chi tiết, như các biên và sự thay đổi cường độ, để tạo ra bản đồ gradient. Cuối cùng, mạng tinh chỉnh kết hợp thông tin từ mạng toàn cục và gradient để cải thiện độ chính xác của bản đồ độ sâu, tạo ra kết quả hoàn thiện hơn. Kiến trúc này kết hợp các đặc trưng toàn cục và cục bộ, mang lại khả năng ước lượng độ sâu chính xác trong các tình huống phức tạp.

Phân loại mạng ước lượng độ sâu

- Ước lượng độ sâu từ ảnh đơn (Monocular Depth Estimation)

Dự đoán bản đồ chiều sâu chỉ từ một ảnh RGB đầu vào.

+ Nguyên lý: Mạng học các đặc trưng phức tạp trong ảnh như phối cảnh, kích thước tương đối, và ngữ cảnh để suy ra khoảng cách tương đối đến từng điểm ảnh.
+ Ưu điểm: Chỉ cần một camera RGB thông thường, dễ áp dụng trong các hệ thống di động.
+ Hạn chế: Chỉ ước lượng tương đối, không tuyệt đối nếu không có thông tin chuẩn (ground truth), kết quả phụ thuộc nhiều vào chất lượng dữ liệu huấn luyện.

- Ước lượng độ sâu từ ảnh đôi (Stereo Depth Estimation)

Sử dụng hai ảnh chụp từ hai camera hoặc hai góc nhìn khác nhau để tính toán độ sâu dựa trên sự chênh lệch (disparity) giữa các điểm tương ứng.

+ Nguyên lý: Tam giác hóa (Triangulation) được sử dụng để tính khoảng cách từ camera dựa vào chênh lệch giữa các điểm tương ứng trên hai ảnh.
+ Ưu điểm: Độ sâu tuyệt đối và chính xác hơn so với ảnh đơn. không phụ thuộc quá nhiều vào dữ liệu huấn luyện.
+ Hạn chế: Yêu cầu hệ thống camera stereo được căn chỉnh chính xác, khó áp dụng trong trường hợp có nhiều vật che khuất.

- Ước lượng độ sâu từ video (Structure from Motion)

Sử dụng chuỗi ảnh hoặc video (nhiều khung hình từ một camera duy nhất) để tính toán độ sâu thông qua chuyển động của camera.

+ Nguyên lý: Từ sự thay đổi vị trí của các đối tượng trong các khung hình, mạng học cách suy ra độ sâu thông qua cấu trúc chuyển động.

+ Ưu điểm: Không yêu cầu dữ liệu stereo, kết hợp hiệu quả thông tin thời gian (temporal information).
+ Hạn chế: Cần thông tin về chuyển động của camera hoặc đối tượng, khó hoạt động trong môi trường tĩnh hoặc phức tạp.

Nguyên lý hoạt động của mạng ước lượng độ sâu

Đầu vào: Một ảnh RGB đơn hoặc nhiều ảnh từ góc nhìn khác nhau.

Xử lý bởi mạng: Mạng CNN hoặc transformer trích xuất đặc trưng không gian và ngữ cảnh. Các mô hình tích hợp khả năng nhận biết ngữ cảnh, phối cảnh, và các yếu tố hình học để dự đoán độ sâu cho từng điểm ảnh.

Đầu ra: Bản đồ chiều sâu (depth map), trong đó mỗi giá trị tại một điểm ảnh tương ứng với khoảng cách tương đối hoặc tuyệt đối từ camera đến điểm đó.

1.3.2. Phương pháp hình học (Geometric Methods)

Dựa trên các nguyên lý hình học và phối cảnh để tính toán khoảng cách.

Tam giác hóa (Triangulation)

- Nguyên lý: Sử dụng hai camera hoặc hai góc nhìn khác nhau (hệ stereo camera) để tìm sự chênh lệch vị trí của cùng một điểm (disparity).
- Công thức:

$$Z = \frac{f.B}{disparity}$$

Trong đó:

Z: Khoảng cách từ camera đến điểm ảnh.

f: Tiêu cự của camera.

B: Khoảng cách giữa hai camera (baseline).

$disparity$: Sự chênh lệch giữa hai vị trí điểm ảnh.

Tỷ lệ đối tượng (Object Scaling)

- Nguyên lý: Nếu biết kích thước thực tế của một đối tượng (ví dụ: bàn tay), khoảng cách có thể được suy ra từ kích thước của đối tượng trong ảnh.
- Ứng dụng: Áp dụng khi đối tượng có kích thước cố định (ví dụ: bàn tay có chiều dài trung bình).