Depth Estimation from Single Image Using CNN-Residual Network

1st Tăng Hoài Duy
Trường Đại Học Công Nghiệp Thành Phố Hồ Chí Minh
Thành Phố Hồ Chí Minh, Việt Nam
tanghoaiduy360@gmail.com

2nd Phạm Hà Văn Đông
Trường Đại Học Công Nghiệp Thành Phố Hồ Chí Minh
Thành Phố Hồ Chí Minh, Việt Nam
vandong875@gmail.com

I. TÓM TẮT

Trong project này, chúng tôi giải quyết vấn đề ước tính độ sâu từ 1 ảnh. Chúng tôi sử dụng một kiến trúc hoàn toàn phù hợp, lần đầu tiên trích xuất tính năng hình ảnh bằng mạng ResNet-50 được đào tạo trước. Chúng tôi thực hiện việc học chuyển giao bằng cách thay thế lớp FC của ResNet-50 bằng các khối upsampling để khôi phục kích thước của bản đồ độ sâu. Khối upsampling kết hợp khái niệm học tập còn lại. Mạng CNN-Residual này có thể được đào tạo từ đầu đến cuối và chạy thời gian thực trên các hình ảnh có đủ sức mạnh tính toán.

Chúng tôi chứng minh rằng phương pháp lấy mẫu mạng CNN-Residual của chúng tôi mang lại kết quả tốt hơn so vớ Fully Connected, vì nó tránh được overfit. Chúng tôi cũng tiến hành xây dựng mô hình của mình với mạng CNN thuần túy và minh họa hiệu quả của việc học tập chậm hơn. Chúng tôi cũng cho thấy ảnh hưởng của các loss function khác nhau trong quá trình đào tạo. Kết quả được hiển thị bằng cách so sánh trực quan những số liệu.

II. GIỚI THIỆU

Là một vấn đề cơ bản trong thị giác máy tính, ước tính độ sâu cho thấy các mối quan hệ hình học trong một ảnh. Các mối quan hệ này giúp cung cấp các biểu diễn phong phú hơn về các đối tượng và môi trường của chúng, thường dẫn đến những cải tiến trong các nhiệm vụ nhận dạng hiện có, cũng như cho phép nhiều ứng dụng khác nhau như 3D modeling, physics and support models [28], autonomous driving, video surveillance, robotics [4].

Trong khi ước tính độ sâu dựa trên stereo images hoặc chuyển động đã được khám phá rộng rãi, ước tính độ sâu từ monocular image thường phát sinh trong thực tế, chẳng hạn như hiểu rõ hơn về nhiều hình ảnh được phân phối trên web và các phương tiện truyền thông xã hội, danh sách bất động sản, v.v., bao gồm cả ví dụ trong nhà và ngoài trời. Do đó, việc phát triển một hệ thống thị giác máy tính có khả năng ước tính bản đồ độ sâu bằng cách khai thác các tín hiệu monocular không chỉ là một nhiệm vụ đầy thách thức mà còn là một nhiệm vụ cần thiết trong các tình huống không có cảm biến độ sâu trực tiếp. Hơn nữa, thông tin độ sâu được biết đến để cải thiện nhiều tác vụ thị giác máy tính liên quan đến ảnh RGB, chẳng hạn như such as in recognition [22] and semantic segmentation [2].

Ước tính độ sâu từ monocular imaget là một vấn đề khó, vì một hình ảnh RGB được chụp có thể tương ứng với vô số cảnh trong thế giới thực và không thể thu được tín hiệu hình ảnh đáng tin cậy nào. Một số công trình đã cố gắng giải quyết vấn đề này. Phương pháp Structure-from-Motion tận dụng chuyển động của máy ảnh để ước tính tư thế máy ảnh thông qua các khoảng thời gian khác nhau và lần lượt, ước tính độ sâu thông qua phép đo tam giác từ các cặp góc nhìn liên tiếp. Các tác phẩm khác sử dụng variations in illumination [34] and focus [31] làm giả định.

Gần đây, Mạng Nơ-ron Tích Chập (CNN) đã được sử dụng để tìm hiểu mối quan hệ giữa các pixel màu và độ sâu. Chúng tôi đã triển khai một kiến trúc CNN có thể đào tạo đầu cuối kết hợp với mạng dư để tìm hiểu ánh xạ giữa cường độ pixel hình ảnh màu với depth map tương ứng.

III. CÔNG VIỆC LIÊN QUAN

A. *Phương Pháp Cổ Điển*

Trong bài toán single-view depth estimation, hầu hết các tác phẩm dựa vào chuyển động của máy ảnh Structure-from-Motion method [21]), sự thay đổi về độ chiếu sáng (Shape-from-Shading [34]) hoặc sự thay đổi trong tiêu điểm (Shapre-from-Defocus [31]).

Nếu không có thông tin như vậy, ước tính độ sâu hình ảnh RGB đơn lẻ cũng đã được nghiên cứu. Các phương pháp cổ điển dựa trên các giả định mạnh mẽ về hình học cảnh, dựa trên các tính năng được làm thủ công và các mô hình đồ họa xác suất khai thác sự liên kết ngang của hình ảnh hoặc thông tin hình học khác. Ví dụ, Saxena và cộng sự. [26] độ sâu dự đoán từ một tập hợp các tính năng hình ảnh sử dụng hồi quy tuyến tính và MRF, và sau đó mở rộng công việc của chúng vào hệ thống Make3D để tạo mô hình 3D [27]. Tuy nhiên, hệ thống dựa vào căn chỉnh theo chiều ngang của hình ảnh và bị ảnh hưởng trong các cài đặt ít được kiểm soát hơn. Lấy cảm hứng từ công việc này, Liu và cộng sự. [15] kết hợp nhiệm vụ phân đoạn ngữ nghĩa với ước lượng chiều sâu, trong đó các nhãn dự đoán được sử dụng như các ràng buộc bổ sung để tạo thuận lợi cho nhiệm vụ tối ưu hóa, Ladicky và cộng sự. [10] thay vào đó cùng dự đoán nhãn và độ sâu trong cách tiếp cận phân loại. Hoiem và cộng sự. [6] không dự đoán độ sâu một cách rõ ràng, mà thay vào đó, phân loại các vùng hình ảnh thành các cấu trúc hình học và sau đó tạo ra một mô hình 3D đơn giản của hiện trường .

B. *Phương pháp ánh xạ dựa trên tính năng*

Loại công việc liên quan thứ hai thực hiện đối sánh dựa trên tính năng giữa hình ảnh RGB nhất định và hình ảnh của kho lưu trữ RGB-D để tìm các vùng lân cận gần nhất, các bản đối chiếu độ sâu được truy xuất sau đó được làm cong và kết hợp để tạo ra bản đồ độ sâu cuối cùng. Karsch và cộng sự. [7] thực hiện cong vênh bằng cách sử dụng SIFT Flow [16], theo sau là sơ đồ tối ưu hóa toàn cục, trong khi Konrad et al. [8] tính toán giá trị trung bình trên các bản đồ độ sâu đã truy xuất, sau đó là lọc song phương để làm mịn. Thay vì làm cong các ứng viên, Liu et al. [19], xây dựng bài toán tối ưu hóa dưới dạng Trường ngẫu nhiên có điều kiện (CRF) với các thế biến liên tục và rời rạc. Đáng chú ý, những cách tiếp cận này dựa trên giả định rằng sự tương đồng giữa các vùng trong hình ảnh RGB cũng ngụ ý các dấu hiệu độ sâu tương tự.

C. *Dựa trên phương pháp CNN*

Gần đây, các phương pháp ước tính độ sâu dựa trên CNN bắt đầu chiếm ưu thế. Vì nhiệm vụ có liên quan chặt chẽ đến việc gắn nhãn ngữ nghĩa, nên hầu hết các công trình đã được xây dựng dựa trên các kiến trúc thành công nhất của Thử thách nhận dạng hình ảnh quy mô lớn ImageNet (ILSVRC) [25], thường khởi tạo mạng của họ bằng AlexNet [9] hoặc VGG sâu hơn [30] . Eigen và cộng sự. [3] là những người đầu tiên sử dụng CNN để ước tính độ sâu hình ảnh đơn lẻ. Các tác giả đã giải quyết nhiệm vụ bằng cách sử dụng hai ngăn xếp mạng sâu. Mạng đầu tiên đưa ra dự đoán độ sâu chung toàn cầu cho toàn bộ hình ảnh và mạng thứ hai tinh chỉnh dự đoán này cục bộ. Ý tưởng này sau đó được mở rộng trong [2], nơi ba ngăn xếp CNN được sử dụng để dự đoán thêm các tiêu chuẩn bề mặt và nhãn cùng với độ sâu. Không giống như các cấu trúc học sâu được sử dụng trong [3, 2], Roy et al. [24] kết hợp CNN với rừng hồi quy [14], sử dụng các kiến trúc rất nông tại mỗi nút cây, do đó hạn chế nhu cầu về dữ liệu lớn.

Một hướng khác để cải thiện chất lượng của các bản đồ độ sâu dự đoán là sử dụng kết hợp CNN và các mô hình đồ họa. Liu và cộng sự. [17] đề xuất tìm hiểu các thế một bậc và từng cặp trong quá trình đào tạo CNN dưới dạng mất trường ngẫu nhiên có điều kiện (CRF) và đạt được kết quả hiện đại mà không cần sử dụng các mồi hình học. Ý tưởng này có ý nghĩa vì các giá trị độ sâu là liên tục [18]. Li và cộng sự. [13] Wang và cộng sự. [32] sử dụng CRF phân cấp để tinh chỉnh các dự đoán CNN thông minh của họ từ superpixel xuống cấp pixel.

D. *Mạng Tích Chập Hoàn Toàn*

Các phương pháp dựa trên học sâu được đề cập ở trên làm tăng đáng kể độ chính xác của bộ dữ liệu điểm chuẩn tiêu chuẩn và là phương pháp tốt nhất hiện nay. Trong khi đó, các nhà nghiên cứu đang cố gắng cải thiện độ chính xác của phương pháp CNN hơn nữa. Nghiên cứu gần đây đã chỉ ra rằng mạng tích chập hoàn toàn (FCN) [20] là một lựa chọn mong muốn cho các bài toán dự đoán dày đặc do khả năng lấy đầu vào có kích thước tùy ý và trả về đầu ra không gian. [1] sử dụng FCN và sử dụng CRF làm hậu xử lý. Bên cạnh các lớp chập cổ điển, [12] sử dụng các lớp chập giãn như một cách hiệu quả để mở rộng trường tiếp nhận của nơ-ron mà không cần tăng các tham số để ước tính độ sâu; [23] sử dụng tích chập chuyển vị để lấy mẫu bản đồ đối tượng và đầu ra để phân đoạn hình ảnh

Laina và cộng sự. [11] đề xuất một mạng được kết nối đầy đủ, loại bỏ các lớp được kết nối đầy đủ và thay thế bằng các khối lấy mẫu dư hiệu quả. Chúng tôi theo dõi sát sao công việc này trong dự án này. Chúng tôi đã hoàn thiện lại kiến trúc trong PyTorch và so sánh hiệu suất của phương pháp này với CNN thuần túy.

IV. PHƯƠNG PHÁP

A. *CNN+FC*

Kiến trúc đầu tiên theo sau công trình ở [3], nơi các tác giả sử dụng mạng CNN thô và tốt để ước tính độ sâu. Về cơ bản, chúng tôi đã thực hiện lại cấu trúc của mạng thô trong bài báo.

Như trong Hình 1 (a), mạng bao gồm hai phần, các lớp chập và các lớp được kết nối đầy đủ. Hình ảnh RGB đầu vào đầu tiên đi qua các lớp tích chập với các bộ lọc 11×11, 5×5, 3×3. batch normalization, các lớp ReLU và các lớp max-pooling 2×2 tuân theo các lớp tích chập. Sau 6 lớp chập, dữ liệu đi qua 2 lớp FC và đầu ra cuối cùng được thay đổi kích thước thành kích thước của bản ground truth depth map. Dropout được sử dụng sau lớp FC đầu tiên để tránh overfitting.

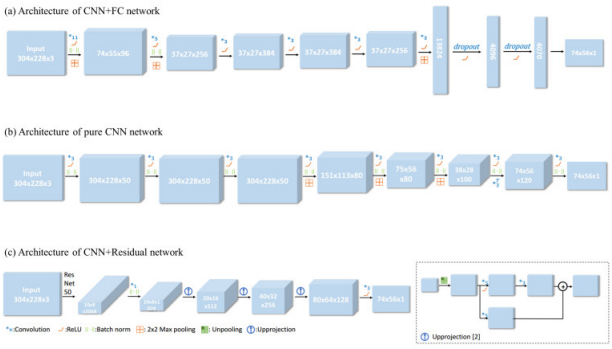
Tổng số parameters của tất cả các lớp tích chập là 27232. Ngược lại, số parameters của hai lớp FC là 7329382. Như được hiển thị trong phần thử nghiệm, mô hình này có thể overfit hàng nghìn hình ảnh, nhưng không đạt được kết quả hợp lý trên validation set. Ngay cả việc thêm các lớp dropout cũng không khắc phục được sự cố.

B. *Pure CNN*

Để khắc phục sự cố overfitting, chúng tôi thay thế các lớp FC bằng các lớp chập. Điều này làm giảm đáng kể số lượng các parameters được sử dụng và khi số lượng các lớp tích chập thích hợp, có thể đạt được hoặc vượt trội so với mạng CNN + FC.

Như trong hình 1 (b), mạng này bao gồm 8 lớp tích chập. Vì các lớp tích chập có tính đến cả global and local information, chúng tôi hy vọng mạng này sẽ có hoạt động tốt hơn. Như đã trình bày trong lớp, các bộ lọc tích chập lớn có thể được thay thế bằng nhiều lớp tích chập hơn với các bộ lọc kích thước nhỏ hơn, điều này làm giảm tổng số parameters để train và có thể thu được kết quả tương tự. Chúng tôi đã thay thế bộ lọc 11×11 và 5×5 bằng bộ lọc 3×3. Mỗi lớp tích chập được theo sau bởi batch normalization để tạo điều kiện thuận lợi cho quá trình train, và sau đó là lớp ReLU. Max-pooling 2×2 được sử dụng để downsample hình ảnh gốc để có được kích thước của depth map.

Chúng tôi cũng đã thử một biến thể của kiến trúc này. Khi chúng tôi đọc qua các bài báo sử dụng phương pháp dựa trên CNN để ước tính độ sâu, chúng tôi nhận thấy rằng hình ảnh gốc thường được lấy mẫu xuống để có đặc điểm chiều cao và chiều rộng nhỏ, sau đó được lấy mẫu lên để có kích thước của bản đồ độ sâu cuối cùng. Vì vậy, chúng tôi đã thay đổi hai lớp cuối cùng của mạng này. Lớp kế cuối được thêm vào lớp



Hình 1. Kiến trúc của ba mô hình được sử dụng trong dự án

max-pooling 2 × 2 và lớp cuối cùng được thay thế bằng một lớp transposed convolution để lấy mẫu đầu vào. Kiến trúc này được gọi là CNN-transpose trong phần thử nghiệm.

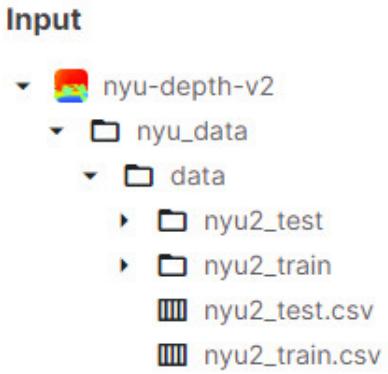
C. *CNN+Residual*

Kiến trúc thứ ba và hứa hẹn nhất của chúng tôi theo sau công trình trong [11]. Bản chất là chúng tôi thực hiện transfer learning với các tính năng hình ảnh được trích xuất và transfer learning không chỉ liên quan đến việc đào tạo lớp FC như chúng ta làm trong các nhiệm vụ phân loại, mà còn tích hợp và upprojection để xây dựng depth map.

Cấu trúc được thể hiện trong Hình 1 (c). Hình ảnh RGB đầu vào được đưa vào làm đầu vào cho mạng ResNet-50 được đào tạo trước. Chúng tôi sử dụng tính năng hình ảnh được trích xuất trước lớp FC, có kích thước 10 × 8 × 2048. Để có được depth map với độ phân giải cao hơn, chúng ta cần thực hiện upsampling. Như được minh họa trong [33], các lớp unpooling làm tăng độ phân giải không gian của bản đồ đối tượng bằng cách thực hiện hoạt động nghịch đảo của việc gộp chung. Khối upprojection được sử dụng trong mạng này bao gồm các lớp unpooling với các lớp chập, được thể hiện ở phía dưới bên phải trong Hình 1. Một tính năng đầu vào có kích thước H × W × C trước tiên đi qua một lớp unpooling, lớp này tăng gấp đôi kích thước đối tượng lên 2H × 2W × C bằng cách ánh xạ từng mục nhập vào góc trên bên trái của kernel 2 × 2. Các mục mở rộng được điền bằng các số không hoặc giá trị trung bình của hai nearest neighbor của nó. Dữ liệu mở rộng sau đó được đưa vào lớp tích chập 5 × 5, được đảm bảo áp dụng cho nhiều hơn một phần tử khác 0 tại mỗi vị trí. Tiếp theo là kích hoạt ReLU và tích chập 3 × 3. Ngoài kiến trúc up convolution này, a projection connection được xây dựng từ feature map có độ phân giải thấp hơn đến block output. Để giữ kích thước thích hợp, up convolution (unpooling và tích chập 5 × 5) cũng được thêm vào nhánh này.

V. BỘ DỮ LIỆU

Chúng tôi sử dụng Bộ dữ liệu NYU Depth V2 [29] cho nhiệm vụ này. Tập dữ liệu này bao gồm các cảnh trong nhà 4K, được quay dưới dạng chuỗi video bằng máy ảnh Kinect của Microsoft. Vì máy ảnh độ sâu và RGB hoạt động ở các tốc độ khung hình thay đổi khác nhau, chúng tôi liên kết từng hình ảnh độ sâu với hình ảnh RGB gần nhất của nó theo thời gian và loại bỏ các khung hình trong đó một hình ảnh RGB được liên kết với nhiều độ sâu. Chúng tôi sử dụng các hình chiếu của máy ảnh được cung cấp cùng với tập dữ liệu để căn chỉnh RGB và các cặp độ sâu; các pixel không có giá trị độ sâu bị thiếu và bị che đi. Để loại bỏ bất kỳ vùng không hợp lệ nào gây ra bởi cửa sổ, cửa mở và bề mặt đặc biệt, chúng tôi cũng che bớt độ sâu bằng mức tối thiểu hoặc tối đa được ghi lại cho mỗi hình ảnh.



VI. KINH NGHIỆM

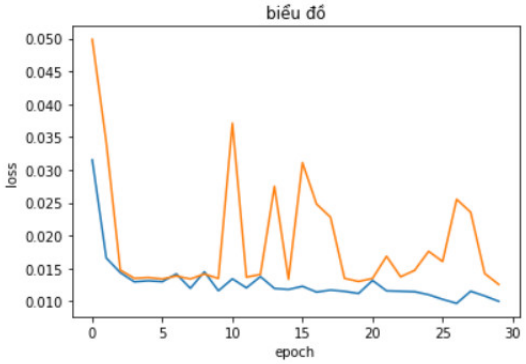
Trong phần này, chúng tôi đưa ra các kết quả định tính của các mô hình bao gồm: CNN, CNN+FC, CNN+Residual .Chúng tôi đánh giá hiệu suất bằng hàm loss function. Tất cả các thử nghiệm được thực hiện trong PyTorch.

A. *Trực quan hóa đầu ra độ sâu*

*a) **CNN+FC**:* Chúng tôi thiết lập một phiên bản đơn giản của kiến trúc CNN + FC và bị overfit trên một tập dữ liệu nhỏ. Trong quá trình triển khai của chúng tôi, phần thô của kiến trúc CNN trong [3] được sử dụng và batch normalization được thêm vào ngay sau mỗi lớp tích chập. Dropout cũng được thêm vào sau lớp FC với xác suất 0,5. Hàm MSE được sử dụng làm loss function của chúng tôi trên cơ sở mỗi pixel. Adam được chọn làm trình tối ưu hóa của chúng tôi, với learning rate được đặt là 1e 4. L2 regularization được thêm vào với sự weight decay là 1e - 4.

50688 hình ảnh được cung cấp cho quá trình train của chúng tôi và 654 hình ảnh được sử dụng để làm validation. Kích thước batch size được đặt thành 32 trong quá trình đào tạo. Kết quả sơ bộ được thể hiện ở 2. Bên trái là loss over time. Từ biểu đồ này, chúng tôi rõ ràng đang bị overfitting dữ liệu của mình:training loss tiếp tục giảm trong khi validation loss giảm đến điểm 1300 sau đó ngừng giảm. Ở bên phải là một số ví dụ từ các bộ đào tạo của chúng tôi. Hàng đầu tiên là hình ảnh đầu vào. Hàng thứ hai là ground truth. Hàng thứ ba là dự đoán về dữ liệu đào tạo của chúng tôi, một lần

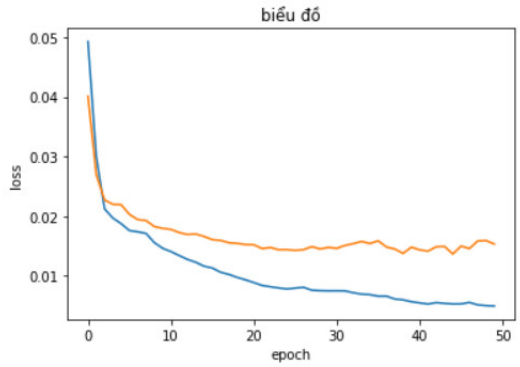
nữa, là kết quả của việc overfitting dữ liệu của chúng tôi.



*b) **Pure CNN**:* Mạng tích chập hoàn toàn này bao gồm 8 lớp tích chập, batch normalization và kích hoạt ReLU theo sau mỗi lớp tích chập. Chúng tôi đã xóa các lớp FC trong mô hình này để tránh overfitting. Mặc dù chúng tôi giảm số lượng parameters để đào tạo đáng kể, việc sử dụng bộ nhớ của các lớp tích chập nhiều hơn và chúng tôi đã giảm kích thước batch size xuống còn 32 để tránh lỗi 'hết bộ nhớ'.Learning rate được đặt thành 1e - 4. Chúng tôi đã sử dụng trình tối ưu hóa Adam trong tác vụ này.

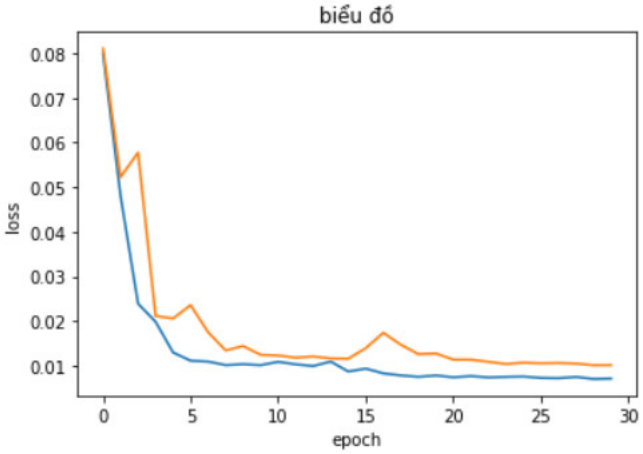
Kết quả của mạng CNN thuần túy và mạng CNN-transposed được thể hiện trong Hình 3. Chúng được đào tạo trên 50688 hình ảnh trong 20 epoch. Khi chúng tôi cố gắng đào tạo với hàng chục nghìn hình ảnh,average training loss rất khó hội tụ, nhưng training và validation loss giảm theo cùng một tốc độ.

The CNN-transposed sử dụng transpose convolution để lấy mẫu và thu được bản đồ độ sâu. Như chúng ta có thể thấy trong hình, cả hai phương pháp đều đưa ra dự đoán độ sâu hợp lý về mặt trực quan, nhưng vẫn tồn tại các hiện vật. Bản đồ độ sâu của mô hình tích chập chuyển vị được lập thành chuỗi. Điều này ngụ ý rằng phương pháp lấy mẫu này không phải là tối ưu. Trong mô hình tích chập chỉ với các lớp chập lấy mẫu, các bản đồ độ sâu dự đoán là mượt mà, nhưng chúng ta có thể thấy rằng các bản đồ độ sâu giống với ảnh gốc hơn là ảnh độ sâu. Nhiều chu kỳ đào tạo hơn có thể giải quyết vấn đề



này.

*c) **CNN+Residual**:* Mô hình CNN + Residual là kiến trúc được [11] đề xuất sử dụng ResNet50 [5] mà không có lớp FC cuối cùng và lớp tổng hợp làm trình trích xuất tính năng, sau đó sử dụng các khối upprojection để lấy mẫu tính năng được trích xuất. Do quy mô của mạng, việc đào tạo mạng này mất nhiều thời gian. Chúng tôi đã đào tạo mạng này trên một tập dữ liệu nhỏ hơn với 500 hình ảnh. Chúng tôi đã thay đổi lớp không chia sẻ trong upprojection và so sánh kết quả trong 5.4. Chúng tôi cũng đã kiểm tra ảnh hưởng của việc sử dụng các thông số được đào tạo trước trong 5.5.



B. *Đánh giá chỉ số*

Đối với hình ảnh có giá trị độ sâu ground truth y và giá trị dự đoán $\hat{y}$, chúng tôi sử dụng ba số liệu khác nhau để định lượng hiệu suất của kiến trúc mạng khác nhau: phần trăm pixel relative error $t = \max\left\{\frac{y}{\hat{y}}, \frac{\hat{y}}{y}\right\}$ nhỏ hơn 1,25. Abs rel diff $\frac{|y-\hat{y}|}{y}$ và RMSE $\sqrt{\frac{1}{n}(\hat{y}-y)^2}$ Chúng tôi so sánh hiệu suất trong 1. CNN + ResNet mang lại hiệu suất tốt nhất. Điều này có thể là do việc trích xuất tính năng tốt được thực hiện bởi ResNet50 được đào tạo trước.

C. *Loss functions: MSE*

Đối với một hình ảnh gồm n pixel với độ sâu thực y và độ sâu dự đoán $\hat{y}$, loss fuctions $B(\hat{y}, y)$ được xác định theo:

MSE: $B(\hat{y}, y) = \frac{1}{n}(\hat{y}-y)^2$

D. *So sánh các phương pháp Unpooling khác nhau*

Trong mô hình CNN + Residual, chúng tôi đã so sánh hai phương pháp unpooling khác nhau trong khối upprojection. Một là sử dụng số 0 để điền vào các mục trống, một là sử dụng giá trị trung bình của hai nearest neighbor để điền vào mục trống (average unpooling). Chúng tôi so sánh hai phương pháp bằng cách huấn luyện mô hình trên một tập dữ liệu nhỏ gồm 25 hình ảnh với 120 epoch. Kết quả được thể hiện trong Bảng 3 và Hình 4

	$t < 1.25$	Abs rel diff	RMSE
Normal Unpooling	0.0537	1.28e6	6.18
Average Unpooling	0.0557	1.27e6	6.37

Table 3. Metric evaluation of different unpooling methods

Từ Hình 4, chúng ta có thể thấy rõ rằng các bản đồ độ sâu được tạo ra của các phương pháp unpooling thông thường có nhiều dạng lưới có khả năng là do phần khác nhau của số không nhận được trong các bộ lọc cuối CNN, nơi đầu ra của average unpooling trơn tru hơn nhiều. Hiệu suất của hai phương pháp unpooling khác nhau là tương tự nhau. Có lẽ

nhiều epoch đào tạo hơn sẽ khiến chúng có sự khác biệt lớn hơn.

E. Ảnh hưởng của Transfer Learning

Có các tham số được đào tạo trước của CNN + Residual trên tập dữ liệu NYU, cho phép transfer learning. Để kiểm tra mức độ ảnh hưởng của các parameters ban đầu, chúng tôi đã đào tạo mô hình CNN + Residual với hai parameters ban đầu khác nhau. Đối với cả hai mô hình, chúng tôi đã sử dụng resNet50 được đào tạo trước [5]. Sau đó, đối với các khối upprojection, chúng tôi khởi tạo một khối với các số ngẫu nhiên nhỏ và khối còn lại với các tham số được đào tạo trước. Chúng tôi đã đào tạo hai mô hình trên tập dữ liệu gồm 500 hình ảnh với 20 epoch. Kết quả được thể hiện trong Bảng 4 và Hình 5.

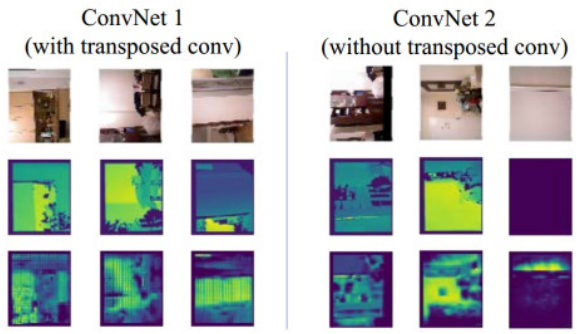


Figure 3. Training results of CNN network

	$t < 1.25$	Abs rel diff	RMSE
CNN+FC	0.307	3.4e5	1.12
CNN	0.195	1.2e6	1.22
CNN+trans	0.143	3.2e5	1.23
CNN+Res	0.634	2.35e4	1.74

Table 1. Metric evaluation results

	$t < 1.25$	Abs rel diff	RMSE
Random Initialization	0.634	2.35e4	1.74
Transfer Learning	0.729	1.43e6	0.734

Table 4. Metric evaluation of different initialization

Từ Hình 5, chúng ta có thể thấy rằng, đối với cùng một số epoch, mô hình khởi tạo ngẫu nhiên hoạt động kém hơn nhiều so với mô hình transfer learning. Có mẫu lưới rõ ràng trong mẫu được khởi tạo ngẫu nhiên. Việc đánh giá thước đo của mô hình transfer learning cũng tốt hơn. Những kết quả này cho thấy lợi ích của việc sử dụng các mô hình được đào tạo trước và thực hiện transfer learning.

VII. KẾT LUẬN

Dự đoán độ sâu bằng cách sử dụng monocular image đóng một vai trò thiết yếu trong nhiều ứng dụng thực tế và là một thách thức vì sự mơ hồ vốn có. Trong dự án này, chúng tôi tiếp cận vấn đề này bằng cách sử dụng CNN và so sánh hiệu suất của các kiến trúc CNN khác nhau trên NYU Depth Dataset V2. CNN với lớp được FC như lớp được sử dụng trong [3] rất mạnh mẽ nhưng có thể dễ dàng overfitting trên tập dữ liệu vì số lượng lớn các parameters trong lớp FC. Điều này thúc đẩy chúng tôi chỉ sử dụng các lớp phức hợp và xếp chồng nhiều lớp hơn để tăng trường tiếp nhận. Kiến trúc này mang lại kết quả có thể chấp nhận được trong khi giảm một số lượng lớn các tham số mô hình. Chúng tôi cũng thử [18] là một kiến trúc CNN sử dụng transfer learning trên ResNet [5] và có thể nhận được kết quả hợp lý trên validation set.

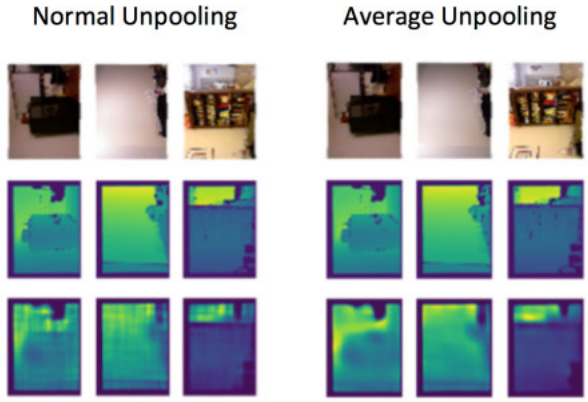


Figure 4. Visualization of different unpooling methods

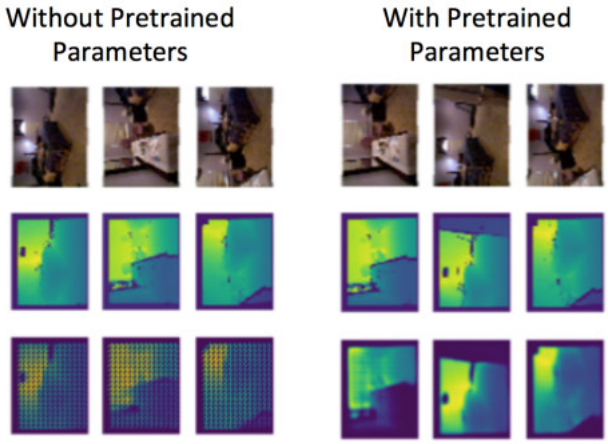


Figure 5. Visualization of different initialization

References

[1] Y. Cao, Z. Wu, and C. Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *arXiv preprint arXiv:1605.02305*, 2016.

[2] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[3] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[4] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun. Learning long range vision for autonomous off-road driving. *Journal of Field Robotics*, 26(2):120–144, 2009.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual

learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM transactions on graphics (TOG)*, 24(3):577–584, 2005.

[7] K. Karsch, C. Liu, and S. Kang. Depth extraction from video using non-parametric sampling. *Computer Vision–ECCV 2012*, pages 775–788, 2012.

[8] J. Konrad, M. Wang, and P. Ishwar. 2d-to-3d image conversion by learning depth from examples. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 16–22. IEEE, 2012.Imagenet [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton.

classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[11] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[12] B. Li, Y. Dai, H. Chen, and M. He. Single image depth estimation by dilated deep residual convolutional neural network and soft-weight-sum inference. *arXiv preprint arXiv:1705.00534*, 2017.

[13] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.

[14] A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[15] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–1260. IEEE, 2010.

[16] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2011.

[17] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.

[18] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2016.

[19] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723, 2014.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[21] G. Qian and R. Chellappa. Structure from motion using sequential monte carlo methods. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 614–621. IEEE, 2001.

[22] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2759–2766. IEEE, 2012.

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing andComputer-Assisted Intervention*, pages 234–241. Springer, 2015.

[24] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514, 2016.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein,et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252,2015.

[26] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

[27] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824– 840, 2009.

[28] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. *Computer Vision–ECCV 2012*, pages 746–760, 2012.

[29] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. *Computer Vision–ECCV 2012*, pages 746–760, 2012.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] S. Suwajanakorn, C. Hernandez, and S. M. Seitz. Depth from focus with your mobile phone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni tion*, pages 3497–3506, 2015.

[32] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L.Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015.

[33] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2018–2025. IEEE, 2011.

[34] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from shading: a survey. *IEEE transactions on pattern analysisand machine intelligence*, 21(8):690–706, 1999.