Journal of Mining and Earth Sciences

Website: https://jmes.humg.edu.vn

Evaluation of DeepLabV3+ with ResNet backbone for building segmentation using UAV images

Dung Trung Pham *, Hung Minh Truong, Phuong Nam Thi Doan, Huong Thu Thi Ta, Ha Thi Nguyen, Men Thi Nguyen

Hanoi University of Mining and Geology, Hanoi, Vietnam

ARTICLE INFO

Article history:
Received 17th Feb. 2025
Revised 02nd May 2025
Accepted 17th May 2025

Keywords:
Building segmentation,
Deep learning,
DeepLabV3+,
ResNet.
UAV.

ABSTRACT

Building segmentation using remote sensing, aerial, and UAV images with deep learning has gained significant attention. Buildings are crucial for urban development, management, and population estimation. Therefore, the automatic extraction of buildings from UAV images is essential for both research and practical applications. This paper presents a building dataset comprising 6,500 image samples, each measuring 512 x 512 pixels, derived from high-resolution UAV images taken with diverse building characteristics in various regions of Vietnam. The study evaluates the effectiveness of building extraction from UAV images using the DeepLabV3+ model with ResNet as the backbone of our dataset. The results indicate that the accuracy for predicting buildings reaches an Intersection over Union (IoU) of 0.774 when employing the ResNet101 backbone. However, this accuracy is significantly influenced by the architectural characteristics and spatial distribution of the buildings. In newly developed urban and suburban areas, the IoU metrics for predicted buildings can reach 0.874 and 0.857, respectively. In contrast, the accuracy declines in industrial zones and older urban areas, with IoU values of 0.762 and 0.673, respectively. This study has practical applications for urban management, development, and the construction of smart cities in our country.

*Corresponding author
E - mail: phamtrungdung@humg.edu.vn
DOI: 10.46326/JMES.2025.66(3).02

Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất
Trang điện tử: https://tapchi.humg.edu.vn

Đánh giá mô hình DeepLabV3+ sử dụng Backbone ResNet để trích xuất tòa nhà từ ảnh UAV

Phạm Trung Dũng *, Trương Minh Hùng, Đoàn Thị Nam Phương, Tạ Thị Thu Hường, Nguyễn Thị Hà, Nguyễn Thị Mến

Trường Đại học Mỏ - Địa chất, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 17/02/2025
Sửa xong 02/5/2025
Chấp nhận đăng 17/5/2025

Từ khóa:
DeepLabV3+,
Mô hình học sâu,
ResNet,
Trích xuất tòa nhà,
UAV.

TÓM TẮT

Trích xuất dữ liệu từ ảnh viễn thám, ảnh hàng không và ảnh UAV sử dụng mạng học sâu đang là hướng nghiên cứu thu hút được nhiều sự quan tâm. Tòa nhà là thông tin trung tâm trong quá trình phát triển và quản lý đô thị cũng như các vấn đề dân số và môi trường. Do đó, việc tự động trích xuất tòa nhà trên tư liệu ảnh là vấn đề được đặt ra cho cả nghiên cứu và trong thực tiễn sản xuất. Bài báo trình bày kết quả trích xuất tòa nhà từ ảnh UAV sử dụng mạng học sâu DeepLabV3+ với (backbone) là cấu trúc mạng phần dư (ResNet) trên bộ mẫu dữ liệu của nhóm nghiên cứu xây dựng. Bộ mẫu dữ liệu tòa nhà gồm 6500 mẫu ảnh có kích cỡ 512 x 512 pixels được xây dựng từ ảnh UAV độ phân giải cao dựa trên sự thay đổi về kiến trúc, hình dạng và phân bố của tòa nhà ở một số tỉnh, thành phố của nước ta. Kết quả chỉ ra rằng độ chính xác dự đoán tòa nhà tính theo chỉ số IoU (tỉ lệ diện tích vùng giao trên vùng hợp) đạt mức 0,774 với backbone ResNet101. Tuy nhiên độ chính xác dự đoán tòa nhà từ mô hình chịu ảnh hưởng lớn bởi đặc điểm kiến trúc và phân bố của tòa nhà. Đối với khu vực đô thị mới và khu vực ngoại ô, độ chính xác dự đoán tòa nhà có thể đạt IoU = 0,874 và 0,857. Tuy nhiên độ chính xác này chỉ đạt IoU = 0,762 và 0,673 đối với khu công nghiệp và khu đô thị cũ. Kết quả của nghiên cứu cho phép ứng dụng mô hình mạng học sâu DeepLabV3+ trong trích xuất dữ liệu tòa nhà phục vụ công tác quản lý và phát triển đô thị cũng như các vấn đề dân số và môi trường ở nước ta.

*Tác giả liên hệ
E - mail: phamtrungdung@humg.edu.vn
DOI: 10.46326/JMES.2025.66(3).02

1. Mở đầu

Thị giác máy tính (computer vision) đã có những bước phát triển vượt bậc khi áp dụng mô hình mạng học sâu (deep learning) trong khoảng hơn 20 năm trở lại đây. Mô hình mạng học sâu tích chập CNN (convolutional neural network) hay DCNN (deep CNN) là mô hình phổ biến được ứng dụng rộng rãi để trích xuất đối tượng trên ảnh và đã thu được nhiều kết quả đầy hứa hẹn (Zhao và nnk., 2024; Bhatt và nnk., 2021; Khan và nnk., 2018). Thành công mang tính đột phá của mạng học sâu trong lĩnh vực thị giác máy tính phải kể tới nghiên cứu về phân loại chữ viết bằng tay của LeCun với mạng học sâu gồm bảy lớp (LeCun và nnk., 1989). Kể từ đó các mạng học sâu hiệu quả hơn trong lĩnh vực thị giác máy tính được ra đời như AlexNet (Krizhevsky và nnk., 2012), GoogLeNet (Szegedy và nnk., 2015), VGGNet (Simonyan, 2014), ResNet (He và nnk., 2016),... Cùng với sự phát triển của các mạng học sâu là các bộ dữ liệu lớn để đào tạo và kiểm chứng khả năng của mô hình mạng học sâu như ImageNet (Deng và nnk., 2009), PASCAL VOC 2012 (Everingham và nnk., 2010), ADE20K (Zhou và nnk., 2019), và Cityscapes (Cordts và nnk., 2016).

Ứng dụng thị giác máy tính trong lĩnh vực trắc địa - bản đồ, viễn thám, và hệ thông tin địa lý để trích xuất các đối tượng như tòa nhà, cây trồng, tuyến đường,... từ ảnh vệ tinh, ảnh hàng không và ảnh UAV đã trở thành xu hướng nghiên cứu được thu hút được nhiều sự quan tâm (Hu và nnk., 2021; Feng và nnk., 2020). Tòa nhà là thông tin trung tâm của đô thị để quản lý phát triển đô thị và xây dựng thành phố thông minh (Wang và nnk., 2022; Huang và nnk., 2022; Xu và nnk., 2018), ước tính dân số (Feng và nnk., 2020), quan trắc thảm họa thiên tai (Long và nnk., 2021; Al Shafian & Hu, 2024; Li & Guo, 2020),... Nhiều mô hình được phát triển dựa trên nguyên tắc của mạng tích chập CNN cho phép trích xuất tòa nhà đạt hiệu suất ngày càng cao. Một số mạng học sâu cơ bản được sử dụng trong trích xuất tòa nhà và các đối tượng trên ảnh hàng không, ảnh viễn thám, ảnh UAV được biết đến gồm mạng Unet (Ronneberger và nnk., 2015), ResNet (He và nnk., 2016), SegNet, Pyramid Scene Parsing Network (PSPNet) (He và nnk., 2015), DeepLab (Chen và nnk., 2017a).

Song song với việc phát triển các mô hình học sâu, các bộ dữ liệu mẫu tòa nhà được xây dựng từ ảnh viễn thám, ảnh hàng không, ảnh UAV phong phú về đặc điểm nhà với độ phân giải cao cũng được giới thiệu. Một số bộ mẫu dữ liệu tiêu biểu có thể kể đến gồm: bộ dữ liệu WHU (Ji và nnk., 2018) bao gồm 220 nghìn tòa nhà trên khu vực có diện tích 450 km^2 tại Christchurch, New ZeaLand có độ phân giải lớn 7,5 cm. Ngoài ra, Mnih (2013) giới thiệu bộ mẫu gồm 151 ảnh có kích thước 1500 x 1500 pixels tại Massachusetts (Hoa Kỳ). Thêm vào đó, International Society for Photogrammetry and Remote Sensing- ISPRS là bộ dữ liệu mẫu tại Vaihingen và Postdam (Đức) (Rottensteiner và nnk., 2012) với độ phân giải cao. Trong đó, bộ mẫu Vaihingen có độ phân giải 9 cm và cỡ ảnh trung bình là 2100 x 2100 pixels, trong khi bộ mẫu Postdam có độ phân giải là 5 cm với cỡ ảnh 6000 x 6000 pixels. Bên cạnh đó phải kể đến bộ mẫu dữ liệu được lấy mẫu tại 10 thành phố trên thế giới với độ phân giải cao 0,3 m thu thập từ ảnh hàng không (Inria Aerial Image Labeling Dataset-INRIA) (Maggiori và nnk., 2017). Bộ dữ liệu này gồm có 360 ảnh màu (RGB) có kích thước 5000 x 5000 pixels.

Sử dụng các mô hình học sâu cải tiến cùng với các bộ dữ liệu mẫu đa dạng đã cho phép trích xuất tòa nhà với những kết quả khả quan. Trong đó, mạng DeepLab là một trong những mạng cải tiến sử dụng nhiều ưu điểm của các mạng học sâu đã giới thiệu trong CNN nên phù hợp cho bài toán trích xuất tòa nhà từ ảnh. Một số kết quả tiêu biểu sử dụng mạng DeepLab trong trích xuất tòa nhà trên ảnh vệ tinh, ảnh hàng không, và ảnh UAV có thể tóm tắt như sau:

Atik và nnk. (2022) đã khảo sát độ chính xác của mạng DeepLabV3+ trong trích xuất toà nhà với các backbone ResNet18, ResNet50, Xception và MobileNetv2 trên một số bộ mẫu dữ liệu chuẩn. Kết quả chỉ ra rằng, độ chính xác sử dụng backbone ResNet50 đạt độ chính xác cao nhất với IoU=0,771 và theo sau là ResNet18 IoU=0,766 với bộ dữ liệu WHUBED. Tương tự như vậy, đối với bộ dữ liệu Massachusetts, độ chính xác đạt giá trị cao nhất IoU=0,612 với backbone ResNet50 và tiếp theo IoU=0,546 với backbone ResNet18. Hơn nữa, trong nghiên cứu (Chen và nnk., 2017b), độ chính xác DeepLabV3 đạt IoU = 0,870 trên tập dữ liệu WHU và đạt IoU = 0,940 trên tập Vaihingen, IoU = 0,794 với tập Postdam trên bộ dữ liệu ISPRS. Ngoài ra, Wang và nnk. (2024) đã đề xuất mô hình MST-DeepLabv3+ bằng cách thay đổi backbone

MobileNetV2 cho Xception đồng thời bổ sung mô đun cổng tập trung (attention mechanism) SENet để tăng độ chính xác phân đoạn hình ảnh. Độ chính xác đạt được của DeepLabV3+ và mô hình tác giả đề xuất lần lượt là mIoU=0,686 và 0,825 (trong đó mIoU thể hiện giá trị trung bình giữa lớp nhà và không phải là nhà) trên bộ mẫu ISPRS. Bên cạnh đó, Li and Zhao (2022) đã sử dụng thêm mô đun cổng tập trung để tối ưu hóa nhánh giải mã (decoder) trong mạng DeepLabV3+ giúp cải thiện độ chính xác phân đoạn hình ảnh. Ngoài ra trong nghiên cứu đó tác giả còn khảo sát việc thay đổi backbone Xception trong phần mã hóa (encoder) bằng cách thay đổi một số lớp trong cấu trúc mạng lưới. Kết quả thực nghiệm trên bộ dữ liệu INRIA chỉ ra rằng, độ chính xác của mạng DeepLabV3+ (IoU = 0,733) được cải thiện 0,6% (IoU = 0,739) khi áp dụng việc chỉnh sửa encoder, và tăng 1,6% khi áp dụng cổng tập trung (IoU = 0,749). Hơn nữa, ý tưởng của mạng DeepLabV3+ còn được sử dụng với cổng tập trung đường bao (attention boundary) để phân tách các đối tượng nhỏ trên ảnh và phân biệt rõ các đối tượng như tuyến đường, cây và bóng đổ của tòa nhà (Xu & Wang, 2024).

Sử dụng mô hình học sâu trong thị giác máy tính nói chung và trích xuất tòa nhà từ tư liệu ảnh nói riêng thường phải tiến hành công tác xây dựng bộ mẫu dữ liệu trên ảnh và đào tạo mô hình từ đầu trên bộ mẫu đã xây dựng. Ưu điểm của đào tạo mô hình từ đầu cho phép kiểm soát các tham số trong quá trình đào tạo mô hình. Điều này giúp tối ưu hóa bộ siêu tham số (hyperparameter) đối với những đối tượng đào tạo cụ thể. Ngoài ra không phải tìm kiếm các mô hình phù hợp với đối tượng nghiên cứu đã được đào tạo từ trước. Tuy nhiên, đào tạo mô hình từ đầu sẽ gây tốn thời gian và tài nguyên của máy tính phục vụ công tác đào tạo mô hình. Nhiều trường hợp tài nguyên máy tính không đáp ứng được yêu cầu hoặc việc xây dựng bộ mẫu cho quá trình đào tạo là không khả thi khi đối tượng nghiên cứu quá đa dạng và phức tạp. Do vậy, để giảm thời gian đào tạo và tránh phải xây dựng bộ mẫu dữ liệu quá lớn trong phương pháp học máy, thì học chuyển giao (transfer learning) sử dụng mô hình có sẵn (pre-trained model) là một giải pháp khả thi. Bởi vì mô hình đã được đào tạo sẵn trên các bộ mẫu dữ liệu lớn có thể học được các tính năng chung và khái quát hóa quá trình học của chúng để áp dụng cho các đối tượng nghiên cứu cụ thể. Bên cạnh đó, mô hình được đào tạo sẵn thường được ứng dụng các mô hình có kiến trúc và kỹ thuật mới nhất có khả năng mang lại hiệu suất cao.

Chính vì thế, việc ứng dụng phương pháp học chuyển giao trên bộ mẫu dữ liệu nhỏ là hướng đi khả thi cho vấn đề này. Tuy nhiên, hiện nay ở nước ta còn thiếu có bộ mẫu dữ liệu chuẩn tòa nhà phục vụ trích xuất dữ liệu từ ảnh độ phân giải cao đặc biệt là đối với ảnh UAV. Do đó, bài báo trình bày phương pháp học chuyển giao sử dụng mô hình DeepLabV3+ với backbone Resnet để tiến hành đào tạo trên bộ mẫu tòa nhà do nhóm nghiên cứu xây dựng. Những đóng góp chính của bài báo bao gồm:

Xây dựng một bộ mẫu dữ liệu nhỏ các tòa nhà phục vụ đào tạo mô hình học sâu cho phép trích xuất tòa nhà từ ảnh UAV phục vụ công tác thành lập bản đồ, quy hoạch đô thị, đánh giá biến động dân cư,... phù hợp với đặc điểm nhà ở nước ta.

Tiến hành đánh giá độ chính xác mô hình phân đoạn hình ảnh DeepLabV3+ với các backbone ResNet có độ sâu khác nhau gồm ResNet18, ResNet34, ResNet50, và ResNet101 trên bộ mẫu dữ liệu của nhóm nghiên cứu xây dựng.

Ngoài ra, bài báo đã huấn luyện được mô hình trên bộ mẫu dữ liệu do nhóm nghiên cứu xây dựng từ đó cho phép dự đoán tòa nhà phù hợp với đặc điểm, hình dạng, cấu trúc của đô thị ở nước ta. Ảnh hưởng của các yếu tố này đến độ chính xác dự đoán tòa nhà cũng được đánh giá theo bốn khu vực bao gồm khu đô thị mới, khu đô thị cũ, khu công nghiệp và khu vực ngoại ô.

Trong phần tiếp theo bài báo trình bày phương pháp nghiên cứu ở phần 2. Phần 3 giới thiệu bộ mẫu dữ liệu nhỏ tòa nhà được thành lập từ ảnh UAV ở một số khu vực ở Việt Nam. Kết quả thực nghiệm và thảo luận sẽ được trình bày trong phần 4. Cuối cùng là một số kết luận, những hạn chế và những hướng nghiên cứu tiềm năng.

2. Phương pháp nghiên cứu

Trích xuất tòa nhà trên dữ liệu ảnh là một trong số những bài toán phổ biến của lĩnh vực thị giác máy tính còn được gọi là phân đoạn ngữ nghĩa (Semantic segmentation). Khác với bài toán xác định đối tượng (Object detection) là tạo khung hình bao quanh đối tượng cần xác định, phân đoạn ngữ nghĩa xác định đối tượng thông qua nhãn

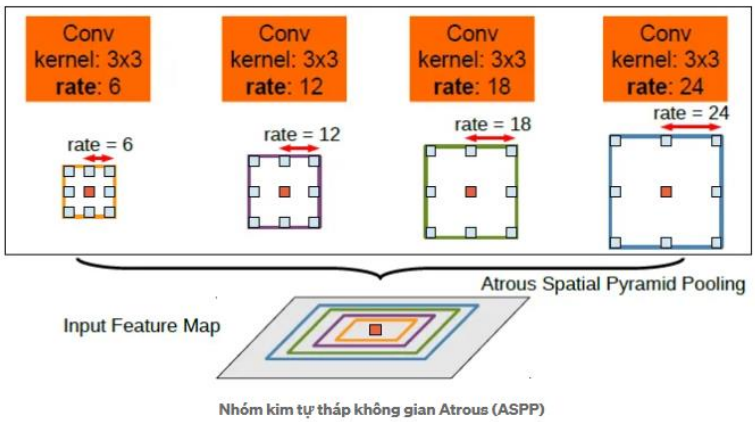
trên từng pixel của ảnh. Hiện nay, mạng học sâu tích chập CNN là mạng phổ biến nhất trong lĩnh vực thị giác máy tính và được áp dụng cho bài toán phân đoạn hình ảnh. Các mạng CNN được sử dụng phổ biến hiện nay cho tác vụ phân đoạn hình ảnh có thể kể đến gồm: VGG, UNet, ResNet, SegNet, DeepLab,... (Luo và nnk., 2021). Trong phần này sẽ tóm tắt nội dung về kiến trúc cơ bản của mạng DeepLab và ý tưởng cơ bản của cấu trúc mạng ResNet.

2.1. Cấu trúc mô hình DeepLab

DeepLab là mô hình phân đoạn ngữ nghĩa được phát triển bởi Google Research và được biết đến với khả năng nắm bắt các chi tiết trên ảnh và phân loại đối tượng dựa trên pixel ảnh. Cải tiến chính của mô hình này là sử dụng tích chập giãn nở (atrous convolution hoặc dilation convolution). Phiên bản đầu tiên (V1) của DeepLab được phát triển dựa trên cấu trúc mạng VGG-16 (gồm 16 lớp) bởi nhóm Visual Geometry Group tại Đại học Oxford (Yu và nnk., 2016). Tuy nhiên có thay thế lớp kết nối đầy đủ (Fully connected layer) cuối cùng bởi lớp tích chập giãn nở.

Đối với các bài toán phân đoạn ngữ nghĩa, số lượng tham số sử dụng thường rất lớn so với các bài toán phân loại khác, vì thế tốc độ tính toán thường sẽ chậm. Do đó, cải tiến quan trọng nhất trong mạng DeepLab là sử dụng tích chập giãn nở để cho phép tăng số lượng tham số mà không làm giảm tốc độ tính toán. Tích chập giãn nở được phát triển từ mạng tích chập chuẩn trong đó bộ lọc được chỉnh sửa để tăng độ lớn của trường tiếp nhận (receptive field) mà không làm tăng số lượng tham số. Tích chập giãn nở hoạt động theo nguyên tắc tạo ra các khoảng trống 'atrous' còn được biết là tỉ lệ giãn nở. Tỉ lệ giãn nở (rate) được ký hiệu (r) để kiểm soát khoảng cách giữa các giá trị trong bộ lọc. Tỉ lệ giãn nở càng cao thể hiện số khoảng trống càng lớn và ngược lại, tỉ lệ r =1 tức bộ lọc chuẩn (không có khoảng trống). Do đó, bằng cách tạo khoảng trống trong bộ lọc, tích chập giãn nở giúp không làm tăng tham số hay thời gian tính toán mà còn cho phép nắm bắt được trường thông tin lớn hơn so với tích chập chuẩn. Việc nắm bắt đối tượng xung quanh (hay ngữ cảnh) giúp nó trở nên phù hợp với mạng phân đoạn ngữ nghĩa. Để tránh mất thông tin do khoảng trống của bộ lọc DeepLab phiên bản V2 đã tích hợp nhiều bộ lọc với các giá trị rate khác nhau còn gọi là mô đun kim tự tháp không gian (Atrous Spatial Pyramid Pooling ASPP) (Hình 1). ASPP cũng là cấu trúc nổi bật được sử dụng trong mạng SPPNet (He và nnk., 2015).

Nguyên tắc hoạt động của ASPP là các tích chập giãn nở được vận hành song song với nhiều tỉ lệ giãn nở khác nhau để trích xuất đặc trưng đầu vào của ảnh. Sau đó, hợp nhất tất cả các đầu ra ở các tỉ lệ khác nhau lại. Ưu điểm của ASPP là mỗi phép tích chập với tỉ lệ cụ thể sẽ thu được thông tin ở một tỉ lệ xác định. Trong đó, tỉ lệ thấp sẽ tập trung vào việc thu thập các chi tiết nhỏ (vật thể nhỏ) trong khi tỉ lệ cao sẽ sẽ thu thập thông tin trên khu vực lớn hơn (vật thể lớn). Vì vậy, sử dụng nhiều tỉ lệ giúp mạng có thể nắm bắt tốt hơn các vật thể có kích thước khác nhau trong hình ảnh vì thế giúp cải thiện độ chính xác phân đoạn hình ảnh.



Hình 1. Nhóm kim tự tháp không gian ASPP là sự kết hợp nhiều atrous convolution với các tỉ lệ khác nhau gồm rate=6, 12, 18 và 24 sử dụng trong DeepLab V2 (Chen và nnk., 2017a).

Mô hình DeepLab phiên bản V3 điều chỉnh lại mô đun ASPP trong phiên bản DeepLabV2 để đạt hiệu suất cao hơn. Theo đó ASPP có một kênh riêng biệt cho nhóm hình ảnh tổng thể cho phép nắm bắt được toàn bộ đặc tính tổng thể, sau đó tổng hợp lại với véc tơ đặc điểm đã trích xuất thông qua tích chập ASPP. Tiếp theo, kết quả cho qua lớp tích chập có đầu vào 1x1 để thu được các chi tiết tốt hơn. Một thay đổi khác đối với mô-đun ASPP là sử dụng kỹ thuật BatchNormalization sau tích chập và trước ReLU để giúp mô hình hội tụ nhanh hơn. BatchNormalization là một kỹ thuật chuẩn hóa đặc biệt được giới thiệu từ năm 2015 (Ioffe, 2015). Kỹ thuật này giúp các bản đồ đặc trưng (feature map) đầu ra được chuẩn hóa về mặt phân phối xác suất từ đó giúp mô hình hội tụ nhanh hơn.

Nguyên lý của tích chập giãn nở (Atrous convolution) trong không gian một chiều với đầu vào x và đầu ra y tại phần tử thứ i được định nghĩa theo công thức sau:

$$y[i] = \sum_{k=1}^{K} x[i + rk]w[k] \qquad (1)$$

Trong đó: *w* - bộ lọc tích chập với độ dài *k*.

Ngoài việc sử dụng tích chập giãn nở thay thế cho phép tích chập chuyển vị 'transposed convolution' của các mô hình khác, DeepLab còn sử dụng trường ngẫu nhiên có điều kiện 'Conditional Random Field (CRF)' để giúp cho dự đoán tốt hơn ở lớp cuối. CRF dự đoán nhãn dựa theo mức xác suất theo công thức sau:

$$E(x) = \sum_{i} \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \qquad (2)$$

Trong đó: *x* - nhãn được gán cho các pixels.

Hệ số đầu tiêu trong công thức (2) sử dụng để đo mức độ giống nhau giữa pixel ảnh được phân đoạn và ảnh thật tương ứng của nó. Đây là lớp ban đầu do mạng tích chập học sâu DCNN dự đoán cho mỗi pixel thuộc lớp tương ứng được định nghĩa bởi $\theta_i(x_i) = -logP(x_i)$ với $P(x_i)$ là xác suất tại pixel thứ *i* được tính bởi DCNN sử dụng hàm mất mát (loss function). Hệ số thứ hai trong công thức (2) sử dụng để tìm mối quan hệ giữa hai pixels bất kỳ trong ảnh. Dựa trên nguyên tắc các pixel có đặc điểm tương tự hoặc có khoảng cách lân cận sẽ được ưu tiên gán nhãn giống nhau. Trong đó thành phần $\theta_{ij}(x_i, x_j)$ được định nghĩa bởi:

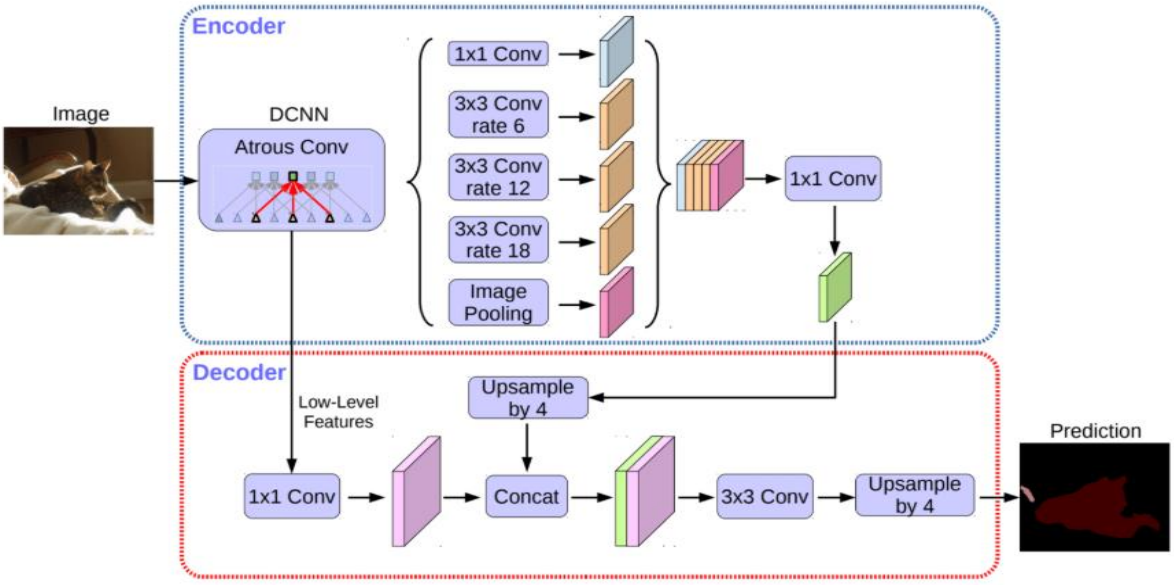
$$\begin{aligned}&\theta_{ij}(x_i, x_j)\\ &= \mu(x_i, x_j)\left[w_1 exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\alpha^2}\right.\right.\\ &\left.-\frac{\|I_i - I_j\|^2}{2\sigma_\beta^2}\right)\\ &\left.+ w_2 exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\gamma^2}\right)\right]\end{aligned} \qquad (3)$$

Trong công thức (3), $\mu(x_i, x_j)$=1 nếu $x_i \neq x_j$, và $\mu(x_i, x_j)$=0 nếu $x_i = x_j$. Các giá trị $w_1, w_2, \sigma, \alpha, \beta, \gamma$ được gọi là các siêu tham số. Khoảng cách giữa p_i và p_j được coi là khoảng cách giữa hai pixels. Ngoài ra, khoảng cách giữa I_i và I_j biểu diễn khoảng cách Euclidean giữa hai giá trị cường độ của pixel tương ứng. Có nghĩa là nếu hai pixel có màu sắc càng khác nhau thì sẽ có khoảng cách lớn hơn. Tuy nhiên, "khoảng cách" có dấu trừ ở phía trước, do đó tác động lớn nhất trong hàm số này chỉ khi cả vị trí lẫn màu sắc tương tự nhau. Do đó, các pixel lân cận có màu tương tự sẽ bị "phạt" rất nặng nếu chúng thuộc các lớp nhãn khác nhau. Nói cách khác, các pixel cạnh nhau và có màu sắc giống nhau sẽ ưu tiên có cùng nhãn. Thành phần thứ 2 trong công thức (3) được thiết kế để kiểm soát độ mịn của hàm số và chỉ xem xét đến khoảng cách giữa các pixel.

Mô hình mạng DeepLabV3+ được phát triển dựa trên phiên bản DeepLabV3 và ý tưởng của mạng Encoder-Decoder (ví dụ mạng UNet (Ronneberger và nnk., 2015)) bằng cách thêm vào một mô đun upsample với output stride (OS=4) trong phần decoder để giảm bộ nhớ của máy tính. Có thể miêu tả khái quát cấu trúc của mạng thông qua bộ mã hóa và bộ giải mã như sau (Hình 2).

Bộ mã hóa (Encoder): Mạng DeepLabV3+ sử dụng lại mô hình DeepLabV3 làm bộ mã hóa (encoder). Đầu ra từ lớp cuối của bộ mã hóa được sử dụng như đầu vào cho cấu trúc mạng Encoder-Decoder ở khối tiếp theo.

Bộ giải mã (Decoder): DeepLabV3 thường trích xuất đặc tính đầu ra sử dụng output stride (OS=16). Các đặc tính được trích xuất này giàu thông tin ngữ nghĩa về mặt hình ảnh. Một bản đồ đặc tính khác được trích xuất từ một lớp nông (Low-level features) gần với đầu vào của backbone với phép tích chập (1x1) để giảm số lượng kênh và chứa thông tin về mặt không gian.



Hình 2. Kiến trúc mô hình mạng DeepLabV3+(Chen và nnk., 2018).

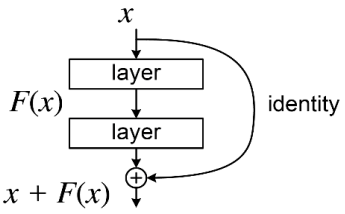
Cùng với đó, đầu ra đầu tiên từ bộ Encoder được lấy mẫu song tuyến với hệ số 4. Tiếp theo, kết quả này cùng với bản đồ đặc trưng chứa thông tin không gian (chứa ranh giới của đối tượng) được gộp lại (Concat) và đưa qua lớp tích chập (3x3) và sau đó được được lấy mẫu song tuyến với hệ số 4 lần thứ hai để đưa ra dự đoán. Kết quả là đối tượng được dự đoán sẽ có đầy đủ đặc tính kèm theo ranh giới của đối tượng.

2.2. Mạng phần dư (ResNet)

Mạng ResNet (Residual Network) được biết đến là mạng học sâu sử dụng cho các bài toán xác định đối tượng và phân đoạn hình ảnh. Hiện nay có rất nhiều biến thể của mạng ResNet với số lớp khác nhau gồm 18 lớp (ResNet18), 34 lớp (ResNet34), 50 lớp (ResNet50), 101 lớp (ResNet101), hoặc 152 lớp (ResNet152).

Trong quá trình phát triển các mạng học sâu, số lớp mạng ngày càng có xu hướng tăng lên (mạng sâu hơn) để mô hình có khả năng tổng quát hóa dữ liệu tốt hơn. Ví dụ: Các mạng học sâu ban đầu như AlexNet với chỉ 5 lớp cho đến các VGGNet có 16, 19 lớp. Vấn đề gặp phải của các mạng khi có nhiều lớp là hiện tượng 'vanishing gradient'. Bản chất của hiện tượng này là giá trị đạo hàm 'độ dốc' rất nhỏ dẫn đến hiệu suất của mạng bị giảm nhanh chóng hoặc thậm chí là bị bão hòa. Chính vì vậy mạng phần dư ra đời để giải quyết vấn đề vanishing gradient dựa trên ý tưởng các kết nối "tắt" (skip connection) để có thể bỏ qua được một hoặc nhiều lớp trong mạng (một cách giảm độ sâu bằng phép nối tắt). Ý tưởng nối tắt (skip connection) mà ResNet sử dụng đươc dựa trên ý tưởng trong mạng Highway Network (Srivastava và nnk., 2015) hay trong Long Short Term Memory LSTM (Hochreiter, 1997). Hình 3 miêu tả sơ đồ phép nối tắt được sử dụng trong mạng ResNet trong đó các lớp mạng (layer) vẫn tương tự như các mạng khác gồm lớp tích chập (convolution), lớp tổng quát hóa (pooling), lớp kích hoạt (activation) và lớp kết nối đầy đủ (fully connected layer). Mũi tên bên phải hình vẽ thể hiện phép nối tắt để bổ sung một giá trị x vào hàm F(x) hay còn gọi là 'phần dư' cũng là tên gọi của mạng 'phần dư- residual network'. Chính vì bổ sung giá trị x mà tránh được giá trị đạo hàm bằng 'không' hay hiện tượng 'vanishing gradient' khi mạng có nhiều lớp.

2.3. Đánh giá kết quả



Hình 3. Sơ đồ nối tắt (skip connection) trong mạng ResNet (He và nnk., 2016).

Trong quá trình xây dựng mô hình phân đoạn hình ảnh, công tác đánh giá kết quả dự đoán của mô hình là bước quan trọng để biết được chất lượng của mô hình. Đánh giá mô hình giúp chúng ta lựa chọn được đúng mô hình và các tham số phù hợp đối với đối tượng nghiên cứu. Trong bài toán phân đoạn hình ảnh chỉ số IoU, Dice là những thước đo được tin dùng để đánh giá độ chính xác mô hình (Jadon, 2020), (Punn và nnk., 2020).

Chỉ số IoU 'hay Jaccard' được tính bởi tỷ lệ giữa diện tích vùng giao và diện tích vùng hợp giữa đối tượng dự đoán 'P-predicted' và đối tượng thực tế 'A-actual' được mô tả như sau:

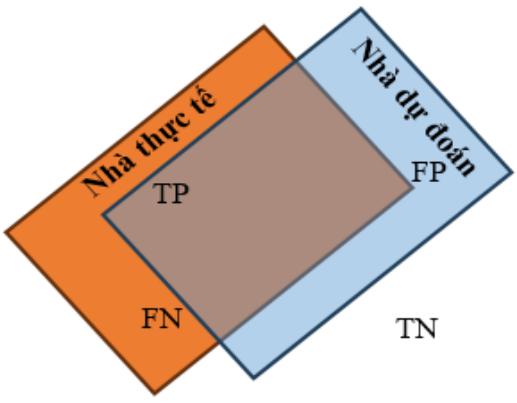
$$IoU = \frac{|P \cap A|}{|P \cup A|} = \frac{TP}{TP + FP + FN}. \quad (4)$$

Giống như Jaccard/IoU, chỉ số Dice (còn biết là Sørensen–Dice) cũng là một thước đo đáng tin cậy trong phân đoạn ngữ nghĩa. Hệ số Dice được tính bằng cách nhân đôi diện tích chồng phủ giữa phân đoạn dự đoán và thực tế, sau đó chia cho tổng số pixel trong cả hai phân đoạn theo công thức:

$$Dice = \frac{2|P \cap A|}{|P| + |A|} = \frac{2TP}{2TP + FP + FN}. \quad (1)$$

Trong công thức (4) và (5) các giá trị TP, FP, TN và FN được định nghĩa như sau (Hình 4).

- True Positive (TP): mô hình dự đoán đối tượng là nhà và thực tế đối tượng đúng là nhà.
- False Positive (FP): mô hình dự đoán đối tượng là nhà nhưng thực tế không phải nhà.
- True Negative (TN): mô hình dự đoán đối tượng không phải là nhà và thực tế đối tượng không phải là nhà.



Hình 4. Bốn khả năng dự đoán tòa nhà.

- False Negative (FN): mô hình dự đoán đối tượng không phải là nhà nhưng thực tế đối tượng lại đúng là nhà.

Hệ số IoU và Dice bằng 1 biểu thị sự chồng phủ giữa nhà dự đoán và nhà thực tế là hoàn hảo hay dự đoán tuyệt đối chính xác. Nếu IoU và Dice bằng 0 có nghĩa là dự đoán không chính xác. Chỉ số này đặc biệt hữu ích trong các tình huống cần xác định chính xác các vùng được phân đoạn mà không quan tâm đến các vùng không được phân đoạn.

3. Dữ liệu

Bộ mẫu dữ liệu tòa nhà phục vụ quá trình đào tạo mô hình phải đảm bảo tính đa dạng và phù hợp với đặc điểm tòa nhà ở nước ta. Đặc điểm của tòa nhà trong trích xuất dữ liệu từ ảnh thường phụ thuộc vào hình dạng, kích thước, kiến trúc, độ cao tòa nhà và màu sắc mái (Li và nnk., 2022). Ngoài ra, các vật xung quanh tòa nhà như tuyến đường, sân và cây cối bao quanh tòa nhà có thể có ảnh hưởng đến quá trình trích xuất dữ liệu tòa nhà. Trong nghiên cứu này, bộ mẫu nhà được tiến hành phân loại tòa nhà theo bốn khu vực gồm: khu đô thị cũ, đô thị mới, khu công nghiệp, khu ngoại ô (Hình 5).

Đối với các tòa nhà trong khu đô thị cũ, đặc điểm chung của chúng là có mật độ xây dựng cao, diện tích nhỏ và thường bị che phủ bởi cây cối xung quanh hoặc bởi bóng của nhà liền kề.

Hình 5. Đặc điểm tòa nhà sử dụng trong bộ mẫu đào tạo mô hình.

Đối với khu đô thị mới, tòa nhà thường được quy hoạch có khoảng không gian xung quanh và mật độ thưa. Trong các khu công nghiệp, các tòa nhà thường có diện tích rất lớn, màu mái thường đồng nhất và ít cây xanh xung quanh. Khu vực ngoại ô có mật độ xây dựng thấp, các tòa nhà thường bố trí có nhà chính và các công trình phụ, có khoảng sân và xen lẫn là cây xanh. Ngoài ra, đặc điểm và mật độ tòa nhà ở nước ta còn thay đổi tùy phụ thuộc vào vị trí địa lý và theo vùng miền (Bắc, Trung, Nam).

Do đó trong nghiên cứu này, bộ mẫu dữ liệu tòa nhà phục vụ cho công tác đào tạo mô hình được thu thập gồm 25 nghìn tòa nhà tại các khu khu đô thị cũ, khu đô thị mới, khu công nghiệp và ngoại ô nhằm tạo nên bộ mẫu nhà có sự đa dạng, đầy đủ và đảm bảo mức độ phức tạp. Ngoài ra, khi xây dựng bộ mẫu dữ liệu phục vụ đào tạo mô hình, vị trí địa lý và yếu tố vùng miền cũng được xem xét đến. Vì vậy, bộ mẫu ảnh UAV được chọn tại nhiều tỉnh, thành phố bao gồm: Bà Rịa-Vũng Tàu, Đồng Nai, Lào Cai, Hà Nội, Hải Phòng, Khánh Hòa, Quảng Ninh để đảm bảo tính phân bố theo vùng miền

Công tác tạo mẫu dữ liệu được thực hiện bằng phần mềm QGIS, theo đó các tòa nhà được số hóa trực tiếp trên ảnh UAV. Để phù hợp với tài nguyên tính toán của máy tính, ảnh được chia thành các mẫu ảnh có kích thước 512x512 pixels với độ phân giải mặt đất của ảnh là 10 cm. Tiếp theo, vì máy tính làm việc trong hệ nhị phân nên các ảnh sau số hóa cũng cần được chuyển sang dạng mã nhị phân. Theo đó các tòa nhà tương ứng với mã 1 và những khu vực không có nhà tương ứng với mã 0 (Hình 6).

Do số lượng bộ mẫu nhỏ nên trong thực nghiệm này chỉ chia thành hai bộ mẫu phục vụ đào tạo mô hình (training dataset) và kiểm tra mô hình (testing dataset). Mục đích để giành tối đa mẫu cho công tác đào tạo mô hình và bỏ qua bước kiểm thử mô hình trên tập validation. Bộ mẫu đào tạo mô hình gồm 5300 mẫu ảnh và 1200 mẫu cho bộ kiểm tra. Cả hai bộ mẫu đều đảm bảo có đủ bốn loại nhà thuộc các khu đô thị mới, đô thị cũ, ngoại ô và khu công nghiệp.

Hình 6. Mẫu ảnh đã chia với kích thước 512x512 pixels (trái) và mặt nạ tương ứng (phải).

4. Kết quả và thảo luận

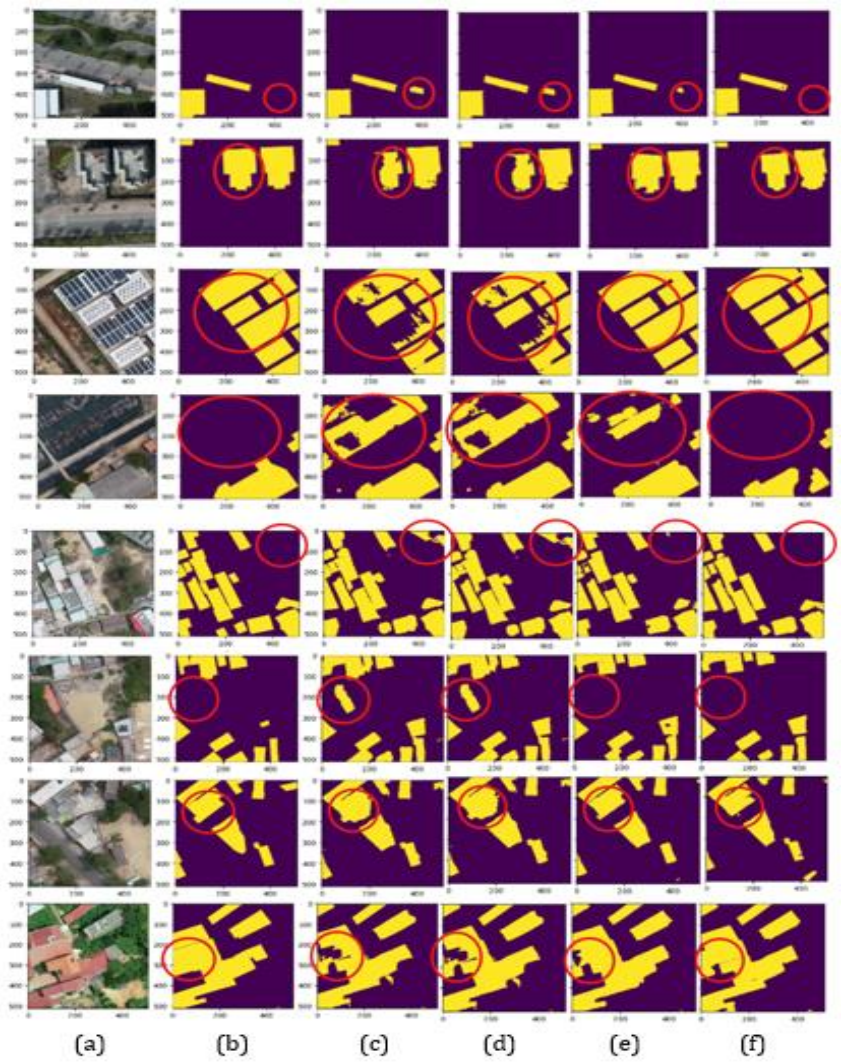
Thực nghiệm đào tạo mô hình được thực hiện trên bộ xử lý đồ họa (GPU) và nền tảng điện toán song song CUDA được cung cấp miễn phí trên Colaboratory. Cấu hình T4 được cung cấp trên Colaboratory là card đồ họa chuyên nghiệp của NVIDIA với cấu hình bộ nhớ GDDR6 16 GB và có 320 lõi tensor giúp tăng tốc trong đào tạo mô hình. Mô hình DeepLabV3+ trong thực nghiệm này được xây dựng trên framework PyTorch 1.11.0. PyTorch hiện nay là một thư viện phổ biến trong nghiên cứu mạng học sâu vì khả năng hỗ trợ của bộ xử lý GPU và có sự hỗ trợ của một cộng đồng nghiên cứu đông đảo. Trong quá trình đào tạo mô hình đã sử dụng (pre-trained model) có bộ trọng số đã được đào tạo trên bộ ImageNet gồm 14 triệu bức ảnh cho phép phân loại 1000 lớp đối tượng.

Quá trình đào tạo mô hình DeepLabV3+ được thực hiện với bốn backbone ResNet có độ sâu 18, 34, 50, 101 lớp tương ứng với các mạng ResNet18, ResNet34, ResNet50, và ResNet101. Vai trò của các backbone này là thay thế nhánh encoder của mô hình DeepLabV3+. Số lượng ảnh cho mỗi lần lấy mẫu (batch_size=16) để phù hợp với bộ nhớ GPU Ram cung cấp bởi Colaboratory. Trong thực nghiệm sử dụng hàm Stochastic Gradient Descent (SGD) để tối ưu hóa hàm mất mát với tốc độ học learning rate là lr=10-4. Tổng số vòng lặp cho cả bốn lần thực nghiệm nêu trên đều giống nhau và bằng 100 epoch.

4.1. Kết quả định tính

Kết quả dự đoán tòa nhà sử dụng các tham số được huấn luyện trên tập kiểm tra (testing dataset) với 1200 mẫu ảnh. Trong Hình 7 từ trái qua phải, cột (a) thể hiện hình ảnh tòa nhà, cột (b) là mặt nạ tòa nhà thực tế, các hình (c), (d), (e), và (f) là các mặt nạ tòa nhà được dự đoán từ mô hình DeepLabV3+ sử dụng backbone ResNet18, ResNet34, ResNet50, và ResNet101, tương ứng.

Kết quả đánh giá định tính thấy rằng mô hình đã phát hiện tương đối đầy đủ các đối tượng là tòa



Hình 7. So sánh kết quả trích xuất tòa nhà sử dụng các backbone ResNet. Trong đó (a)- ảnh mảu RGB, (b) mặt nạ tòa nhà thực, kết quả dự đoán nhà với backbone (c)ResNet18, (d) ResNet34, (e) ResNet50, và (f) ResNet101.

nhà. Sự sai sót trong dự đoán của mô hình thường do một trong các nguyên nhân như: (1) mô hình nhận nhầm các đối tượng có hình dạng giống nhà, (2) không nhận ra các đối tượng nhà có màu sắc mái giống tương tự với màu của đối tượng xung quanh (nhà bị lẫn với đối tượng khác), hoặc (3) tòa nhà bị cây che phủ một phần mái cũng bị mô hình dán sai nhãn.

Cụ thể, mô hình sử dụng các backbone ResNet50 và ResNet101 cột (e) và (f), tương ứng đã trích xuất được các tòa nhà tương đối giống với mặt nạ tòa nhà thực ở cột (b). Ngược lại, mô hình sử dụng backbone ResNet18 và ResNet34 có mặt nạ dự đoán tòa nhà ở cột (c) và (d) có sự khác biệt tương đối lớn so với mặt nạ tòa nhà thực ở cột (b). Điều này thể hiện độ chính xác dự đoán tòa nhà của các backbone là mô hình ResNet50 và ResNet101 tốt hơn so với hai backbone còn lại. Kết quả này có thể được giải thích bởi độ sâu (số lớp) của mô hình resnet được sử dụng. Mô hình có độ sâu lớn hơn cho phép trích xuất tốt hơn các đặc trưng của đối tượng và ngược lại mô hình nông sẽ trích xuất kém hơn.

4.2. Kết quả định lượng

Kết quả dự đoán tòa nhà trên tập testing dataset được tính bởi hai chỉ số IoU và Dice được tổng hợp trong Bảng 1. Độ chính xác có xu hướng tăng với độ sâu của mạng ResNet sử dụng. Đối với mạng ResNet18 giá trị IoU = 0,730 và tăng lên

khoảng 0,3%, 1,0% và 4,4% tương ứng với mạng ResNet34 (IoU=0,733), ResNet50 (IoU=0,740), và ResNet101 (IoU = 0,774). Độ chính xác dự đoán tòa nhà đạt được trên mức 70% đo bởi chỉ số IoU là hoàn toàn phù hợp với nghiên cứu (Atik và nnk., 2022).

Bảng 1. Độ chính xác dự đoán tòa nhà với Deeplabv3+ sử dụng các backbone ResNet trênbộ 1200 mẫu gồm khu đô thị cũ, đô thị mới, khu công nghiệp và ngoại ô (thời gian được tính trên 1 epoch).

Backbone	Tham số	Thời gian	Độ chính xác	
			IoU	Dice
ResNet18	11 Mb	4'18"	0,730	0,822
ResNet34	21 Mb	4'34"	0,733	0,829
ResNet50	23 Mb	4'45"	0,740	0,832
ResNet101	42 Mb	9'08"	0,774	0,843

Tương tự như chỉ IoU, chỉ số Dice cũng ghi nhận xu hướng tăng từ Dice=0,822 khi sử dụng ResNet18 đến Dice = 0,843 khi sử dụng ResNet101. Giá trị Dice tăng khoảng 2% khi sử dụng mạng ResNet có độ sâu lớn hơn từ 18 lớp lên 101 lớp.

Mặc dù độ chính xác dự đoán của mô hình DeeplabV3+ tăng với độ sâu của backbone ResNet sử dụng, tuy nhiên thời gian thực hiện đào tạo mô hình lại tăng lên tỉ lệ thuận với số lượng tham số của mô hình như biểu diễn trong cột 2 và 3 của Bảng 1. Chỉ số về thời gian và độ chính xác của mô hình là hai yếu tố quan trọng trong quyết định lựa chọn mô hình. Tùy thuộc vào mục đích sử dụng mà ưu tiên độ chính xác mô hình (ví dụ trong trường hợp tối ưu hóa mô hình về độ chính xác) hoặc cần mô hình "nhẹ" với thời gian tính toán nhanh. Trong nhiều trường hợp người ta căn cứ vào cả hai chỉ tiêu độ chính xác và thời gian để chọn ra mô hình "cân bằng" giữa độ chính xác và thời gian tính toán.

Từ bộ dữ liệu trên tiến hành đánh giá độ chính xác tòa nhà theo từng khu vực dựa theo đặc điểm kiến trúc và mật độ tòa nhà. Theo đó bộ mẫu dữ liệu gồm 1200 mẫu được chia ra với 940 mẫu khu đô thị cũ, 50 mẫu khu đô thị mới, 160 mẫu khu công nghiệp và 50 mẫu khu vực ngoại ô.

So sánh độ chính xác dự đoán tòa nhà dựa theo các khu đô thị cũ, đô thị mới, khu công nghiệp và ngoại ô được miêu tả trong các Bảng 2÷ 5 tương ứng. Kết quả chỉ ra rằng, độ chính xác dự đoán nhà cũng phụ thuộc vào đặc điểm của tòa nhà tại từng khu vực. Trong đó, độ chính xác dự đoán tòa nhà ở khu đô thị mới đạt độ chính xác cao nhất và theo sau là khu vực ngoại ô với chỉ số tương ứng là IoU=0,874 và 0,857 sử dụng backbone ResNet 101. Kết quả này cũng phù hợp với nghiên cứu của (Chen và nnk., 2017b) khi sử dụng bộ mẫu nhà trong khu quy hoạch. Độ chính xác dự đoán nhà tại khu đô thị cũ đạt độ chính xác thấp nhất với IoU=0,673 với cùng backbone ResNet101. Độ chính xác tính theo chỉ số Dice cũng có xu hướng tương tự như chỉ số IoU đã phân tích ở trên.

Bảng 2. Độ chính xác dự đoán tòa nhà với Deeplabv3+ sử dụng các backbone ResNet trên bộ 940 mẫu tại khu đô thị cũ.

Backbone	Đánh giá độ chính xác	
	IoU	Dice
ResNet18	0,664	0,857
ResNet34	0,664	0,858
ResNet50	0,667	0,861
ResNet101	0,673	0,861

Bảng 3. Độ chính xác dự đoán tòa nhà với Deeplabv3+ sử dụng các backbone ResNet trên bộ 50 mẫu tại khu đô thị mới.

Backbone	Đánh giá độ chính xác	
	IoU	Dice
ResNet18	0,872	0,961
ResNet34	0,871	0,960
ResNet50	0,874	0,962
ResNet101	0,874	0,962

Bảng 4. Độ chính xác dự đoán tòa nhà với Deeplabv3+ sử dụng các backbone ResNet trên bộ 160 mẫu tại khu công nghiệp.

Backbone	Đánh giá độ chính xác	
	IoU	Dice
ResNet18	0,762	0,852
ResNet34	0,701	0,812
ResNet50	0,695	0,809
ResNet101	0,750	0,844

Bảng 5. Độ chính xác dự đoán tòa nhà với Deeplabv3+ sử dụng các backbone ResNet trên bộ 50 mẫu khu vực ngoại ô.

Backbone	Đánh giá độ chính xác	
	IoU	Dice
ResNet18	0,842	0,948
ResNet34	0,842	0,948
ResNet50	0,847	0,950
ResNet101	0,857	0,953

Kết quả này phù hợp với đặc điểm và mức độ phức tạp của các tòa nhà tại bốn khu vực khảo sát. Đối với khu đô thị mới, tòa nhà thường có tính đồng nhất và có khoảng không gian xung quanh tách biệt với các tòa nhà khác nên mô hình có thể dễ dàng dự đoán chính xác. Tương tự như vậy, đối với khu vực ngoại ô, mật độ xây dựng thưa và tòa nhà thường có khoảng không gian xung quanh cũng là những điều kiện thuật lợi để mô hình dự đoán đúng tòa nhà. Ngược lại, khu đô thị cũ có mật độ xây dựng cao, các tòa nhà được xây liền nhau và ít khoảng không gian sẽ khó khăn cho mô hình dự đoán đúng.

Đối với khu công nghiệp, độ chính xác dự đoán tòa nhà cao nhất ở mức (IoU = 0,762 và Dice=0,852) sử dụng backbone ResNet18. Mặc dù đặc điểm nhà của khu công nghiệp khá tương đồng về màu sắc và ít bị lẫn bởi cây cối và vật xung quanh, tuy nhiên độ chính xác dự đoán tòa nhà lại không cao. Kết quả này có thể được giải thích bởi kích thước tòa nhà thường quá lớn so với mẫu ảnh sử dụng để đào tạo mô hình. Trong thực nghiệm này mẫu ảnh có kích thước (512 x 512 pixel) với độ phân giải mặt đất là 10 cm nên trong nhiều trường hợp một mẫu ảnh chỉ có thể biểu diễn được một phần của tòa nhà trên ảnh mà không đầy đủ hình dạng tòa nhà. Giải pháp trong trường hợp tòa nhà lớn và rất lớn có thể sử dụng kích thước ảnh lớn hơn ở mức 1024 x1024 pixel hoặc thậm chí 2048 x 2048 pixels. Bên cạnh đó, độ chính xác dự đoán tòa nhà ở khu công nghiệp có xu hướng không phụ thuộc vào độ sâu của lớp mạng backbone ResNet sử dụng. Kết quả này chưa thể được giải thích trong nghiên cứu này.

5. Kết luận

Trích xuất tòa nhà từ ảnh UAV sử dụng mạng học sâu đã đạt được những kết quả triển vọng. Nghiên cứu này tiến hành đánh giá khả năng trích xuất tòa nhà từ ảnh UAV độ phân giải cao sử dụng mạng DeepLabV3+ với các backbone ResNet khác nhau. Kết quả dự đoán tòa nhà từ bộ mẫu dữ liệu nhỏ được thành lập từ ảnh UAV ở một số khu vực ở nước ta chỉ ra rằng độ chính xác dự đoán tòa nhà có thể đạt được trên 70% tính theo chỉ số IoU và 80% theo chỉ số Dice. Ngoài ra, độ chính xác dự đoán tòa nhà phụ thuộc rất lớn và đặc điểm của tòa nhà, mật độ xây dựng cũng như bối cảnh xung quanh của tòa nhà. Đối với khu vực có mật độ xây dựng thấp, hình dạng tòa nhà đồng nhất, màu sắc mái tương đồng, có khoảng trống xung quanh tòa nhà thì mô hình có xu hướng dự đoán chính xác hơn. Ngược lại, đối với khu vực có mật độ xây dựng cao, hình dạng nhà đa dạng về kích cỡ và màu sắc mái, các đối tượng xung quanh có hình ảnh lẫn với tòa nhà hoặc một phần mái bị che phủ bởi cây mô hình dự đoán tòa nhà thường có độ chính xác thấp.

Mặc dù mô hình đã dự đoán được tòa nhà tương đối chính xác, tuy nhiên vẫn còn hiện tượng dự đoán quá mức (phát hiện sai những vật có hình dạng giống nhà) hoặc hiện tượng dự đoán dưới mức hay dự đoán thiếu (dự đoán được một phần của ngôi nhà) còn biết đến là hiện tượng overfitting. Điều này được lý giải bởi mô hình chưa có khả năng tổng quát tất cả các yếu tố của tòa nhà từ bộ dữ liệu sử dụng đào tạo. Do đó khi dự đoán còn thiếu sót một số đặc tính của tòa nhà trên bộ kiểm tra (testing). Điều này có thể được khắc phục bởi tăng số lượng mẫu đào tạo và giảm sai sót trong quá trình tạo nhãn trong bộ đào tạo.

Ngoài ra, trong giới hạn của nghiên cứu này ảnh hưởng của siêu tham số (hyperparameter) tới độ chính xác dự đoán mô hình chưa được phân tích.

Từ các phân tích về giới hạn của bài báo, các nghiên cứu về chuẩn hóa bộ mẫu và áp dụng các mô hình học máy tiên tiến đóng vai trò quan trọng trong việc nâng cao độ chính xác của dự đoán tòa nhà. Hơn nữa, để cải thiện hiệu quả mô hình, cần tăng cường số lượng và đa dạng hóa các mẫu nhà trong bộ dữ liệu đào tạo. Bên cạnh đó, kích thước ảnh cũng nên được xem xét kỹ lưỡng, đặc biệt đối với các tòa nhà có quy mô lớn và rất lớn. Ngoài ra, việc áp dụng các mô hình học sâu mới với khả năng tối ưu cả về độ chính xác lẫn tốc độ xử lý sẽ là hướng nghiên cứu cần thiết. Cuối cùng, các thuật toán cho phép nâng cao độ chi tiết và chính xác của đường bao tòa nhà cũng nên được nghiên cứu sâu hơn trong các công trình tiếp theo.

Lời cảm ơn

Bài báo này được hoàn thành từ kinh phí hỗ trợ của đề tài B2024-MDA-09.

Đóng góp của tác giả

Phạm Trung Dũng - đóng góp lên ý tưởng, viết chương trình máy tính và viết bản thảo; Trương

Minh Hùng, Đoàn Thị Nam Phương - đánh giá và chỉnh sửa bài viết; Tạ Thị Thu Hường, Nguyễn Thị Hà, Nguyễn Thị Mến - tạo mẫu dữ liệu.

Tài liệu tham khảo

Al Shafian, S., & Hu, D. (2024). Integrating machine learning and remote sensing in disaster management: A decadal review of post-disaster building damage assessment. *Buildings*, 14(8), 2344.

Atik, S. O., Atik, M. E., & Ipbuker, C. (2022). Comparative research on different backbone architectures of DeepLabV3+ for building segmentation. *Journal of Applied Remote Sensing*, 16(2), 024510-024510.

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S.,,... Ghayvat, H. (2021). CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, 10(20), 2470.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834-848.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017b). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv*:1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801-818.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R.,,... Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (CVPR), 2016, pp. 3213-3223.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition,* Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. J. I. j. o. c. v. (2010). The pascal visual object classes (voc) challenge. 88, 303-338. https://doi.org/10.1007/s11263-009-0275-4.

Feng, W., Sui, H., Hua, L., Xu, C., Ma, G., & Huang, W. J. I. J. o. R. S. (2020). Building extraction from VHR remote sensing imagery by combining an improved deep convolutional encoder-decoder architecture and historical land use vector map. *International Journal of Remote Sensing,* 41(17), 6595–6617. https://doi.org /10.1080/01431161.2020.1742944.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

He, K., Zhang, X., Ren, S., Sun, J. J. I. t. o. p. a., & intelligence, m. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(9), 1904-1916.

Hochreiter, S. J. N. C. M.-P. (1997). Long Short-term Memory. *Neural Computation* 8(9), 1735-1780. DOI: 10.1162/neco.1997.9.8. 1735.

Hu, Q., Zhen, L., Mao, Y., Zhou, X., & Zhou, G. J. A. i. C. (2021). Automated building extraction using satellite remote sensing imagery. *Automation in Construction* 123, 103509. https://doi.org /10.1016/j.autcon.2020.103509.

Huang, J., Li, P., Wang, W., & Pei, Y. (2022). Research on Building Extraction method based on Object-oriented and ArcGIS Engine. *2022 3rd International Conference on Geology, Mapping and Remote Sensing (ICGMRS), IEEE.* DOI: 10.1109/ICGMRS55602. 2022.9849324.

Ioffe, S. J. a. p. a. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*, PMLR 37:448-456, 2015.

Jadon, S. (2020). A survey of loss functions for semantic segmentation. *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, IEEE DOI: 10.1109/CIBCB48159.2020.9277638.

Ji, S., Wei, S., Lu, M. J. I. T. o. g., & sensing, r. (2018). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing* 57(1), 574-586. DOI: 10.1109/TGRS.2018.2858817.

Khan, S., Rahmani, H., Shah, S. A. A., Bennamoun, M., Medioni, G., & Dickinson, S. (2018). A guide to convolutional neural networks for computer vision. *Springer Cham.* https://doi.org/10.1007/978-3-031-01821-3.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. J. A. i. n. i. p. s. (2012). Imagenet classification with deep convolutional neural networks. *Publication History* 6(60) 84-90. https://doi.org/10.1145/3065386.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. J. N. c. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1(4), 541-551. DOI: 10.1162/neco.1989.1.4.541.

Li, J., Huang, X., Tu, L., Zhang, T., & Wang, L. (2022). A review of building detection from very high resolution optical remote sensing images. *GIScience & Remote Sensing* 59(1), 1199-1225. https://doi.org/10.1080/15481603.2022.2101727.

Li, W., & Zhao, S. (2022). Semantic segmentation of buildings in high-resolution remote sensing images based on DeepLabV3+ algorithm. In *Journal of Physics: Conference Series* (Vol. 2400, No. 1, p. 012037). IOP Publishing.

Li, Z., & Guo, Y. (2020). Semantic segmentation of landslide images in Nyingchi region based on PSPNet network. *2020 7th International Conference on Information Science and Control Engineering (ICISCE), IEEE*. DOI: 10.1109/ICISCE50968.2020.00256.

Long, L., He, F., & Liu, H. J. T. J. o. S. (2021). The use of remote sensing satellite using deep learning in emergency monitoring of high-level landslides disaster in Jinsha River. *J Supercomput* 77, 8728–8744 (2021). https://doi.org/10.1007/s11227-020-03604-4.

Luo, L., Li, P., & Yan, X. J. E. (2021). Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies* 2021, 14, 7982. https://doi.org/10.3390/en14237982.

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *2017 IEEE International geoscience and remote sensing symposium (IGARSS)*, IEEE DOI: 10.1109/IGARSS.2017.8127684

Mnih, V. (2013). *Machine learning for aerial image labeling*. University of Toronto (Canada). University of Toronto (Canada) ProQuest Dissertations & Theses, 2013. NR96184.

Punn, N. S., Agarwal, S. J. A. T. o. M. C., Communications,, & Applications. (2020). Inception u-net architecture for semantic segmentation to identify nuclei in microscopy cell images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*,16(1), 1-15. https://doi.org/10.1145/3376922

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October* 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer international publishing.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., & Breitkopf, U. (2012). The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3, 1(1), 293-298.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556.

Srivastava, R. K., Greff, K., & Schmidhuber, J. J. a. p. a. (2015). Highway networks. Machine Learning (cs.LG); *Neural and Evolutionary Computing (cs.NE)* https://doi.org/10. 48550/ arXiv.1505.00387.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D.,,... Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (CVPR), 2015, pp. 1-9

Wang, Y., Yang, L., Liu, X., & Yan, P. J. S. R. (2024). An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+. *Sci Rep* 14(1), 9716. https://doi.org/10.1038/s41598-024-60375-1

Wang, Z., Xu, N., Wang, B., Liu, Y., & Zhang, S. (2022). Urban building extraction from high-resolution remote sensing imagery based on multi-scale recurrent conditional generative adversarial network. *GIScience & Remote Sensing* 59(1), 861-884. https://doi.org/10. 1080/15481603.2022.2076382

Xu, S., & Wang, Y. (2024). Fusion of fractal features DeepLabV3+ remote sensing image building segmentation. *2024 43rd Chinese Control Conference (CCC)*, IEEE DOI:10.23919/CCC6 3176.2024.10662351

Xu, Y., Wu, L., Xie, Z., & Chen, Z. J. R. S. (2018). Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens* 10(1), 144. https://doi.org/10.3390/rs10010144.

Yu, W., Yang, K., Bai, Y., Xiao, T., Yao, H., & Rui, Y. (2016). Visualizing and comparing AlexNet and VGG using deconvolutional layers. *Proceedings of the 33 rd International Conference on Machine Learning*, (Vol. 3, pp. 43-76).

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4), 99.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. J. I. J. o. C. V. (2019). Semantic understanding of scenes through the ade20k dataset. 127, 302-321. *Int J Comput Vis* 127, 302–321 (2019). https://doi.org/10. 1007/s11263-018-1140-0.