TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Ứng dụng mạng học sâu tự động trích xuất thông tin công trình xây dựng từ dữ liệu UAV phục vụ công tác quản lý đô thị và hỗ trợ thành lập mô hình 3D LoD1

Nguyễn Hữu Long[1], Phạm Văn Chung[2], Phạm Thị Làn[2], Nguyễn Văn Trung[2,4], Lê Thị Thu Hà[2,4], Đặng Diệu Huệ[3], Phạm Trung Dũng[2]*

[1] Trường Đại học Đồng Tháp; nhlong@dthu.edu.vn
[2] Trường Đại học Mỏ - Địa chất; phamvanchung@humg.edu.vn; nguyenvantrung@humg.edu.vn; lethithuha@humg.edu.vn; phamtrungdung@humg.edu.vn
[3] Đại học xây dựng Hà Nội; huedd@huce.edu.vn
[4] Nhóm nghiên cứu Công nghệ Địa tin học trong Khoa học Trái đất (GES), Trường Đại học Mỏ - Địa chất; nguyenvantrung@humg.edu.vn; lethithuha@humg.edu.vn

*Tác giả liên hệ: phamtrungdung@humg.edu.vn; Tel.: +84–904303904

Ban Biên tập nhận bài: 5/11/2024; Ngày phản biện xong: 25/12/2024; Ngày đăng bài: 25/6/2025

Tóm tắt: Tự động trích xuất đối tượng trên ảnh UAV là quá trình nhận dạng và véc tơ hóa đối tượng trực tiếp từ dữ liệu ảnh. Trong công tác quản lý và phát triển đô thị hiện nay, các công trình xây dựng như tòa nhà và các kiến trúc xây dựng cần được thể hiện trên môi trường đồ họa, lưu trữ và quản lý thống nhất trên hệ thống cơ sở dữ liệu. Bài báo này giới thiệu một phương pháp mới sử dụng công nghệ học sâu để tự động trích xuất mái của các tòa nhà và công trình xây dựng trực tiếp trên ảnh UAV. Dữ liệu mái nhà được trích xuất nhanh chóng được kết hợp với độ cao và các thông tin sẵn có của tòa nhà hoặc công trình xây dựng cho phép thành lập và cập nhật mô hình 3D LoD1. Kết quả trích xuất tòa nhà được thử nghiệm trên mô hình mạng học sâu U-Net tại hai khu vực đô thị mới và cũ đều đạt kết quả mức trên 60% về độ chính xác. Kết quả bài báo hoàn toàn có thể áp dụng trong công tác quản lý xây dựng tại các đô thị lớn có mức tốc độ đô thị hóa cao ở nước ta hiện nay.

Từ khóa: Tự động trích xuất đối tượng; Mạng học sâu; Mô hình 3D LoD1; Quản lý đô thị; UAV.

1. Giới thiệu

Mô hình 3D là dữ liệu mô tả đặc điểm vị trí không gian địa lý và hình dạng, kích thước trong không gian 3 chiều với các thuộc tính của một đối tượng cụ thể trong thế giới thực được hiển thị, truy xuất, xử lý, biên tập, khai thác, sử dụng thông qua các phần mềm chuyên dụng trong môi trường đồ họa, máy vi tính [1]. Tùy thuộc vào mục đích sử dụng mà mô hình 3D công trình nhà cao tầng được biểu diễn gần đúng của thế giới thực trong đó các tính năng của công trình được mô hình hóa ở một cấp độ cụ thể được gọi là LoD (*Level of Detail*) [2]. LoD là thuật ngữ sử dụng để biểu thị mức độ chi tiết trong mô hình 3D cả về mặt hình học và thuộc tính của đối tượng và nó cũng là một trong những thành phần cơ bản trong hệ thống thông tin địa lý (GIS) và mô hình thành phố 3D. LoD càng chi tiết thì mô hình 3D của đối tượng càng được mô tả giống đối tượng đó trong thế giới thực. LoD được chia thành nhiều cấp khác nhau, mỗi cấp độ sẽ thể hiện mức độ chi tiết thông tin và mức độ tin cậy của các thông tin được đưa vào các thành phần mô hình. Đối với chuẩn CityGML, mức độ chi tiết

của mô hình không gian được chia thành 5 mức độ [3], bao gồm LoD0, LoD1, LoD2, LoD3, LoD4. Đối tượng nghiên cứu của mô hình 3D trong nghiên cứu này chỉ tập trung vào LoD1 tức là đảm bảo thể hiện các khối nhà trong không gian 3D một cách đơn giản mà không bao gồm các chi tiết cấu trúc mái, cửa sổ hoặc của chính.

Công tác thành lập mô hình thành phố 3D có thể sử dụng một số phương pháp [4] bao gồm: dựa trên ảnh hàng không lập thể [5], kết hợp ảnh hàng không và bản đồ có sẵn, sử dụng công nghệ thông tin để nội suy dữ liệu không gian địa lý 3D [6]. Phương pháp quét laser [7], phương pháp sử dụng hệ thống bản đồ di động (*Mobile mapping system*) [8]. Bên cạnh đó, còn rất nhiều các phương pháp kết hợp từ những phương pháp khác nhau để có độ chi tiết và độ chính xác cao hơn theo từng nhu cầu cụ thể [9]. Công nghệ LiDAR có thể được sử dụng để bổ sung độ cao cho các cấu trúc 2D của thành phố từ các nguồn dữ liệu khác nhau để xây dựng dữ liệu không gian địa lý 3D [7]. Tuy nhiên công nghệ này còn hạn chế về phân tích không gian do thực hiện trên hai nguồn dữ liệu độc lập mô hình số bề mặt (DTM) và mô hình tòa nhà (DBM). Công nghệ này có thể cải tiến từ ảnh chụp máy bay, hoặc ảnh chụp mặt đất để bổ xung cấu trúc, và hình ảnh các tòa nhà [10]. Xuất phát từ yêu cầu mức độ chi tiết LoD trong mô hình 3D cần thành lập mà chúng ta có thể chọn lựa một hoặc kết hợp những phương pháp nêu trên để thu thập dữ liệu không gian.

Các mô hình tòa nhà, công trình xây dựng 3D đóng vai trò quan trọng trong nhiều ứng dụng bao gồm quy hoạch đô thị và xây dựng thành phố thông minh. Các phương pháp mô hình hóa tòa nhà 3D là sự kết hợp của phần mái công trình và độ cao cũng như các thông tin khác của công trình. Để tiếp cận việc trích xuất tự động các tòa nhà 3D cấp độ thể hiện ở mức độ chi tiết 1 (LoD1), bài báo giới thiệu phương pháp học sâu để trích xuất dữ liệu phần mái tòa nhà và công trình xây dựng. Phương pháp học máy nói chung và học sâu nói riêng nhằm mục đích làm cho hệ thống máy tính học được khả năng giải quyết một nhiệm vụ cụ thể từ dữ liệu đào tạo được cung cấp [11]. Việc trích xuất tòa nhà có thể được coi là bài toán phân loại với nhiều phương pháp tiếp cận khác nhau trong đó phương pháp học máy là phương pháp phổ biến và hiệu quả hiện nay [11]. Trong trường hợp có đủ bộ mẫu dữ liệu chứa đựng những thông tin về các tòa nhà thì phương pháp học máy có thể giải quyết một cách hiệu quả. Các phương pháp học máy có thể được chia thành các phương pháp "học nông" và "học sâu" dựa trên độ sâu của cấu trúc mô hình được sử dụng [12]. Phương pháp học máy truyền thống thường sử dụng các mô hình học nông và các tính năng thủ công. Trước kia, việc trích xuất dữ liệu từ ảnh hàng không, ảnh viễn thám thường sử dụng các phương pháp truyền thống ví dụ công nghệ nhận dạng nhân tạo (artificial recognition) [13]. Đối với công tác trích xuất lớp nhà, đã tồn tại một số phương pháp cho phép trích xuất dữ liệu tòa nhà dựa trên màu sắc mái, hình dạng thiết kế, bóng, mép cạnh của nhà... Một số kỹ thuật tiêu biểu để giải quyết vấn đề trên bao gồm khớp theo mẫu (*template matching*), lý thuyết đồ thị (*graph theory*), rừng cây lựa chọn ngẫu nhiên (*random forests*), máy vector hỗ trợ (support vector machines) dựa trên việc xác định mái tòa nhà [14, 15].

Ngược lại, phương pháp học sâu được đào tạo từ một bộ mẫu rất lớn với khả năng trích xuất một lượng lớn các đặc trưng từ cấu trúc mô hình và tính năng của đối tượng. Các nghiên cứu sử dụng mạng học sâu trích xuất dữ liệu từ ảnh UAV cũng đã đạt được những kết quả khả quan. Trích xuất dữ liệu nhà từ ảnh UAV có độ phân giải cao chủ yếu sử dụng phương pháp học sâu nhờ sự phát triển của thị giác máy tính [16–19]. So với các phương pháp truyền thống, học sâu có lợi thế trong việc tự động trích xuất đặc điểm của các đối tượng trên ảnh. Mạng thần kinh tích chập CNN (*convolutional neural network*) [20] đã phát triển nhanh chóng và được sử dụng rộng rãi trong phân loại ảnh (*semantic segmentation*) [21] từ các đối tượng tự nhiên và phát hiện mục tiêu (*object detection*) [17].

Tại Việt Nam, nghiên cứu về việc trích xuất dữ liệu tòa nhà sử dụng mạng học sâu vẫn còn nhiều hạn chế. Trong nghiên cứu [22] đã sử dụng mạng Mask R-CNN với bộ dữ liệu ảnh vệ tinh có độ phân giải 0,5 m để trích xuất đối tượng nhà ở khu vực nông thôn của nước ta. Ngoài ra, nghiên cứu [23] đã xây dựng bộ dữ liệu cho khu vực quận Cầu giấy, Hà Nội gồm

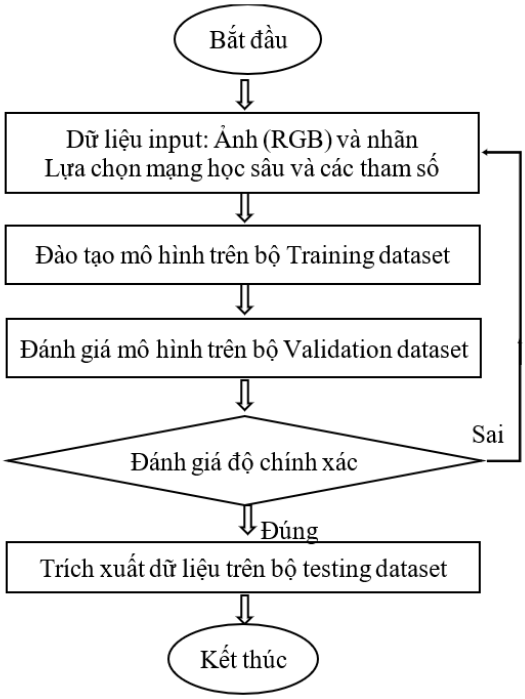
2100 bức ảnh có kích thước 1024×1024 pixel từ Google Earth. Sử dụng mạng U-Net và đánh giá trên tập đào tạo mô hình với độ chính xác tổng thể (*Overall accuracy = 92%*).

Trong nghiên cứu này, mô hình học sâu U-Net được sử dụng để trích xuất tự động mái tòa nhà và các công trình xây dựng trên ảnh UAV tại khu đô thị. Dữ liệu phần mái được trích xuất từ các tòa nhà và công trình là thông tin quan trọng kết hợp với các thông tin khác của tòa nhà để sử dụng xây dựng mô hình 3D LoD1 trong khu vực đô thị. Kết quả nghiên cứu này hoàn toàn có thể áp dụng trong công tác quản lý xây dựng tại các đô thị lớn có mức độ đô thị hóa cao ở nước ta hiện nay.

2. Phương pháp nghiên cứu

2.1. Quy trình trích xuất dữ liệu sử dụng mạng học sâu

Trích xuất dữ liệu tòa nhà trên ảnh UAV sử dụng mạng học sâu đòi hỏi phải tuân thủ theo một quy trình hợp lý. Nhìn chung, có nhiều quy trình khác nhau được đề xuất cho công việc này [24, 25], tuy nhiên, một số bước cơ bản có thể được tóm tắt như trong sơ đồ Hình 1.



Hình 1. Sơ đồ quy trình trích xuất dữ liệu từ ảnh UAV sử dụng mô hình học sâu.

Bước chuẩn bị dữ liệu và thiết kế mô hình là bước đầu tiên và quan trọng trong công tác trích xuất dữ liệu. Trong đó chuẩn bị một bộ dữ liệu phù hợp là một công việc mất nhiều thời gian, công sức và thường chiếm tới 70-80% khối lượng của công việc này. Chi tiết về các bước tạo bộ mẫu dữ liệu cho mô hình học sâu có thể tham khảo trong [26]. Mô hình mạng được lựa chọn thường là các mạng đã thành công trong việc trích xuất hình ảnh trong các bài toán thị giác máy tính. Mạng nơ-ron tích chập CNN là mạng phổ biến và hiệu quả nhất cho đến nay [27].

Sau khi đã chuẩn bị được bộ dữ liệu chuẩn và lựa chọn mô hình mạng học sâu cùng với các giá trị siêu tham số (*hyperparameters*) khởi đầu, công tác đào tạo mô hình có thể được thực hiện. Việc đào tạo mô hình thường được tiến hành với một bộ mẫu lớn về kích thước, đa dạng về chủng loại. Bộ mẫu này thường có kích thước chiếm khoảng 80% số lượng mẫu.

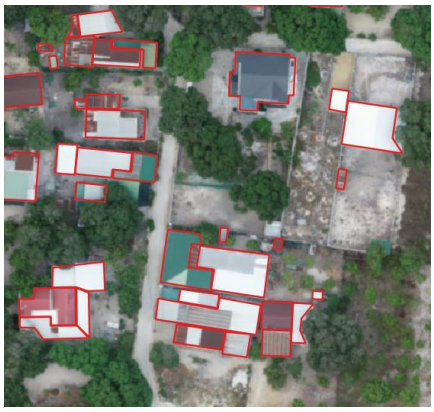
Song song với quá trình đào tạo là quá trình tự đánh giá trên tập validation dataset (còn được gọi là phát triển mô hình). Trong quá trình này cần phải đánh giá độ chính xác của mô hình thông qua các thước đo phù hợp như accuracy, recall, F1-score, IoU,…[28, 29]. Nếu mô hình đảm bảo sự hội tụ thì có thể dừng việc đào tạo và tiến hành thực hiện bước tiếp theo. Nếu mô hình chưa hội tụ, (tức độ chính xác không bảo đảm) thì quay lại kiểm tra lại dữ liệu của bộ mẫu và chọn lựa lại mô hình hoặc cài đặt lại các siêu tham số hyperparameters. Cuối cùng, khi đã có bộ tham số chuẩn chúng ta tiến hành dự đoán trên bộ mẫu dự liệu thực tế testing dataset.

2.2. Dữ liệu

a) Thu thập dữ liệu ảnh: Dữ liệu đầu vào cho mô hình học sâu là các ảnh và nhãn của đối tượng đã được số hóa trực tiếp trên ảnh. Ảnh được sử dụng cho nghiên cứu này được sử dụng là ảnh có độ phân giải cao được thu thập dữ thiết bị bay không người lái UAV. Độ chính xác của ảnh đảm bảo công tác thành lập bản đồ theo các quy phạm hiện hành. Việc thu thập dữ

liệu ảnh UAV và xử lý dữ liệu ra ảnh trực giao (ortho images), mô hình số độ cao (DEM), và dữ liệu đám mây điểm (point clound) có thể tham khảo trong các tài liệu [30–33].

b) Vector hóa tòa nhà trên ảnh UAV: Các tòa nhà sẽ được thực hiện số hóa thủ công trên ảnh UAV sử dụng phần mềm đồ họa AutoCAD hoặc trên QGIS, ArgGIS (Hình 2).



Hình 2. Polygon của đối tượng nhà sau khi số hóa trên ảnh UAV.

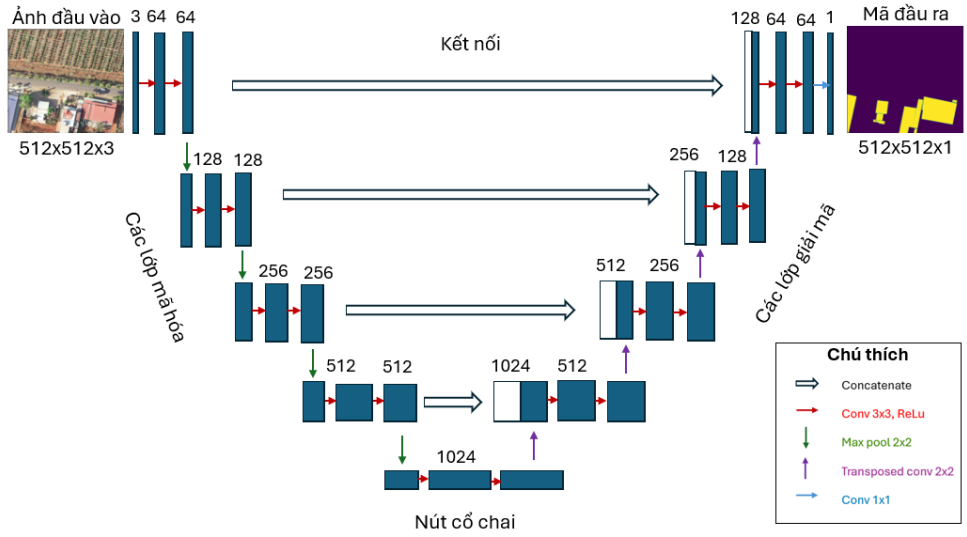
c) Chia mẫu phục vụ mô hình học sâu: Ảnh tòa nhà số hóa trên ảnh nắn có kích thước rất lớn không khả thi cho việc phát triển mô hình học máy. Việc đọc toàn bộ hình ảnh vào bộ nhớ và sử dụng cho việc phát triển mô hình là không khả thi vì đòi hỏi nguồn tài nguyên rất lớn về phần cứng máy tính. Vì thế, để phù hợp với cấu hình phần cứng của máy tính hiện nay, ảnh được chi nhỏ thành các mẫu ảnh (hay tiles). Kích thước của tiles phụ thuộc vào cách thiết kế mô hình và thường có kích thước (256×256), (512×512) hoặc (1024×1024) pixel (Hình 3a). Sau khi đã chia toàn bộ ảnh thành các mẫu nhỏ sẽ thực hiện công việc rater hóa các polygon thành các "mặt nạ - mask" với mức độ xám là 8 bit (Hình 3b).

Hình 3. Các tiles được chia nhỏ từ ảnh UAV và các mặt nạ "mask" tương ứng.

2.3. Xây dựng mô hình mạng học sâu U-Net

Mạng nơ-ron nhân tạo được sử dụng trong trích xuất hình ảnh là mạng tích chập. Mạng nơ-ron tích chập (CNN) là từ viết tắt của cụm *Convolutional Neural Network*. Bản chất của CNN là áp dụng phép tính convolution cho mạng neural network để xử lý các bài toán có số lượng các tham số (*parameters*) vô cùng lớn mà vẫn lấy ra được các đặc trưng của ảnh. Đây là mô hình cơ bản được áp dụng nhiều trong các bài toán học sâu Deep Learning trong lĩnh vực thị giác máy tính hiện nay cho phép người dùng xây dựng những hệ thống phân loại và dự đoán với độ chính xác cao. Trong đó, mạng U-Net là mạng cơ bản ứng dụng nguyên tắc của mạng CNN được sử dụng nhiều hơn trong xử lý ảnh, cụ thể là nhận diện đối tượng trên ảnh UAV, ảnh vệ tinh.

Mạng U-Net [34] lấy tên từ chính cấu trúc hai nhánh hình chữ U của nó. Mạng chữ U bao gồm các lớp tích chập của hai nhánh tương ứng với nhánh mã hóa encoder và nhánh giải mã decoder. Mạng mã hóa có nhiệm vụ phân tích hình ảnh đầu vào để trả lời cầu hỏi đối tượng gì trong ảnh. Đây chính là nhiệm vụ phân loại hình ảnh tương tự như mạng tích chập Convolution neural network (CNN) tuy nhiên đầu ra không phải là nhãn do mạng U-Net không sử dụng lớp fully connected layers. Thay vào đó đầu ra của nhánh mã hóa trong U-Net là các mask có cùng kích thước với hình ảnh đầu vào (Hình 4).



Hình 4. Mạng U-Net với ảnh RGB đầu vào kích thước 512×512×3 và mã hóa đầu ra kích thước 512×512×1.

Các lớp mã hóa (nằm ở phía bên tay trái) chịu trách nhiệm giảm kích thước mẫu và trích xuất đặc điểm để giảm độ phân giải không gian của ảnh trong khi tăng độ sâu của chúng. Điều này cho phép mô hình nắm bắt các biểu diễn trừu tượng của ảnh đầu vào. Đường dẫn mã hóa chứa 5 lớp khối tích chập, trong đó mỗi lớp thực hiện hai phép tích chập 3×3 theo sau là hàm kích hoạt tuyến tính chỉnh lưu (ReLU). Sau đó, kết quả được lấy mẫu xuống với phép gộp tối đa 2×2 với bước nhảy (stride) là 2.

Ngược lại, các lớp giải mã (ở phía bên tay phải) sẽ có nhiệm vụ giải mã dữ liệu và định vị các đặc điểm trong khi vẫn duy trì độ phân giải không gian của ảnh đầu vào. Đường dẫn giải mã hoạt động tương tự như đường dẫn mã hóa, với phép gộp tối đa được thay thế bằng phép tích chập làm tăng gấp đôi chiều rộng và chiều cao của hình ảnh. Các kết nối bỏ qua được sử dụng để giữ lại thông tin không gian bị mất và định vị chính xác các đặc điểm. Hình ảnh được lấy mẫu lên trong đường dẫn mã hóa được nối với bản đồ đặc điểm tương ứng của đường dẫn giải mã.

2.4. Đánh giá độ chính xác mô hình dự đoán

Trong quá trình xây dựng mô hình trong Machine Learning và Deep Learning, một phần không thể thiếu để biết được chất lượng của mô hình đó là đánh giá mô hình. Đánh giá mô hình giúp chúng ta lựa chọn được mô hình mạng học sâu phù hợp đối với bài toán của mình và việc đánh giá mô hình phải dựa trên các "thước đo" phù hợp. Trong bài toán phân loại đối tượng (tòa nhà) trên ảnh ta có thể xem xét việc mô hình phát hiện ra tòa nhà trên ảnh UAV được xác định bởi các thước đo dựa trên ma trận sai số gồm accuracy, precision, recall và F1-score được tính toán như sau [35, 36]:

Accuracy là thước đo mang tính tổng quát được xác định chung cho độ chính xác của cả tòa nhà và độ chính xác của lớp còn lại (tức không phải tòa nhà):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision được định nghĩa là số lượng pixel được phân đoạn chính xác chia cho tổng số pixel mà mô hình dự đoán:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall được tính là số lượng kết quả dương tính thực chia cho tổng số kết quả dương tính thực và kết quả âm tính giả:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

F1-score là giá trị trung hòa giữa precision và recall được tính theo giá trị của precision và recall theo công thức:

$$\text{F1} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{4}$$

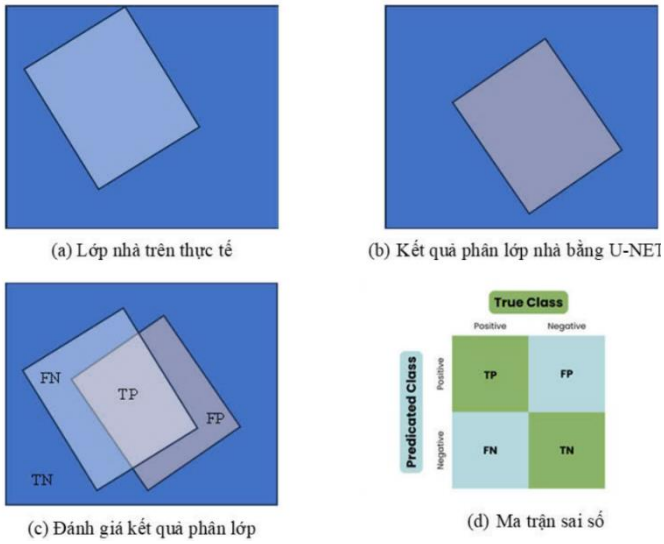
Các giá trị TP, FP, TN và FN được thống kê trong Bảng 1 và thể hiện chi tiết trong Hình 5, có thể được giải thích như sau: True Positive (TP): mô hình dự đoán đối tượng là nhà và đúng là nhà; False Positive (FP): mô hình dự đoán đối tượng là nhà nhưng thực tế không phải nhà; True Negative (TN): mô hình dự đoán tượng không phải là nhà và đúng đối tượng không phải là nhà; False Negative (FN): mô hình nhận diện đối tượng không phải là nhà nhưng thực tế đối tượng lại đúng là nhà.

Bảng 1. Bốn khả năng khi dự đoán tòa nhà trên ảnh.

Dự đoán/Thực tế		**Thực tế**	
		Nhà	**Không phải nhà**
Dự đoán	Nhà	True Positive (TP)	False Positive (FP)
	Không phải nhà	False Negative (FN)	True Negative (TN)

Ngoài bốn thước đo nêu trên, trong bài toán phân đoạn ngữ nghĩa, các chỉ số IoU và Dice là những thước đo được tin dùng để đánh giá độ chính xác mô hình [37, 38].

Chỉ số IoU, còn được gọi là chỉ số Jaccard, đo lường mức độ giao nhau giữa phân đoạn dự đoán và vùng thực. Số liệu này cung cấp một chỉ báo rõ ràng về mức độ phù hợp của phân đoạn dự đoán với phân đoạn thực tế. IoU tính toán tỷ lệ giao nhau và hợp nhau giữa diện tích dự đoán (P) và diện tích phân đoạn thực tế (A) được mô tả như sau:



Hình 5. Bốn khả năng dự đoán nhà trên ảnh.

$$\text{IoU} = \frac{|\text{P} \cap \text{A}|}{|\text{P} \cup \text{A}|} = \frac{|\text{P} \cap \text{A}|}{|\text{P}| + |\text{A}| - |\text{P} \cap \text{A}|} \tag{5}$$

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 chéo giữa phân đoạn dự đoán và thực tế, sau đó chia cho tổng số pixel trong cả hai phân đoạn theo công thức:

$$\text{Dice} = \frac{2|\text{P} \cap \text{A}|}{|\text{P}| + |\text{A}|} \tag{6}$$

Hệ số Dice bằng 1 biểu thị sự chồng chéo hoàn hảo, trong khi 0 biểu thị không có sự chồng phủ. Chỉ số này đặc biệt hữu ích trong các tình huống mà trọng tâm là xác định chính xác các vùng được phân đoạn mà không quan tâm nhiều đến các vùng không được phân đoạn.

3. Kết quả

3.1. Dữ liệu khu vực nghiên cứu

Khu vực nghiên cứu bao gồm những khu đô thị tại các tỉnh, thành Bà Rịa - Vũng Tàu, Khánh Hòa, Quảng Ninh, Hà Nội, Hải Phòng. Đặc điểm của các tòa nhà trong khu vực đô thị có hình dạng và kích thước khác nhau, chủ yếu là hình chữ nhật, một số hình chữ L hoặc hình vuông. Đặc điểm của nhà ở đô thị của nước ta hiện nay thường có mật độ rất dày ở khu vực nội đô (đô thị cũ) và có mật độ vừa phải, được quy hoạch tốt ở những khu đô thị mới. Các tòa nhà ở khu vực đô thị cũ thường có diện tích nhỏ và thường không có ranh giới rõ ràng với các đối tượng xung quanh. Ngược lại, các tòa nhà ở khu vực đô thị mới thường lớn hơn và được bao quanh bởi không gian mở. Các loại tòa nhà có hình dạng, kích thước khác nhau được sử dụng cho tập dữ liệu nhằm mục đích tăng độ phức tạp và tính đa dạng trong quá trình đào tạo mô hình học sâu.

Các mẫu hình ảnh trong tập dữ liệu được chụp bằng UAV có độ phân giải không gian là 10 cm (mỗi pixel tương ứng với diện tích 10×10 cm^2 trên mặt đất). Toàn bộ khu vực nghiên cứu có khoảng 20 nghìn tòa nhà tương ứng với 5000 ô mẫu có kích thước 512×512 pixel tương ứng. Để phục vụ quá trình đào tạo, phát triển và kiểm tra mô hình, bộ dữ liệu được chia thành các tập hợp con bao gồm: tập đào tạo (training dataset), tập phát triển (*validation dataset*) và tập thử nghiệm (*test datasest*) theo tỷ lệ 80%:10%:10%.

(a) **(b)**

Hình 6. Đặc điểm nhà trong khu đô thị mới và cũ: (a) Khu đô thị mới; (b) Khu đô thị cũ.

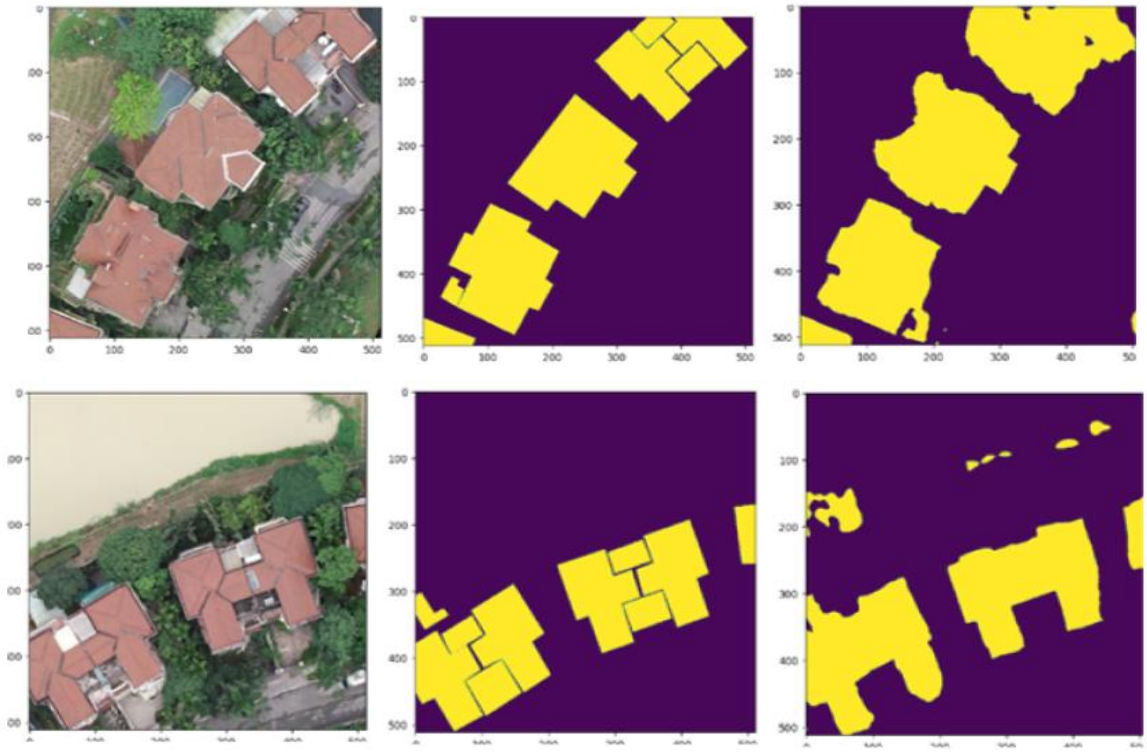
3.2. Cấu hình máy tính và tham số được cài đặt cho mô hình học sâu

Các thực nghiệm được thực hiện bằng cách sử dụng ngôn ngữ PyTorch 1.11.0. Google Collaboratory. Bộ xử lý đồ họa (GPU) và nền tảng điện toán song song CUDA được cung cấp miễn phí trên Collaboratory với thời gian hạn chế. GPU đã giảm đáng kể thời gian đào tạo bằng cách cho phép các phép tính riêng biệt chạy song song. Cấu hình GPU cụ thể được sử dụng là Tesla T4, một card đồ họa chuyên nghiệp của NVIDIA với bộ nhớ GDDR6 16 GB và giao diện bộ nhớ 256 bit. Ngoài ra, nó có 320 lõi tensor giúp tăng tốc độ của các ứng dụng học sâu. Trong thực nghiệm này, kích thước mỗi batch_size ảnh được chọn là 8 để phù hợp với 16G Ram bộ nhớ được cung cấp trên Collaboratory và tổng số vòng lặp được đặt là 100. Hàm Cross Entropy được sử dụng để tính hàm mất mát và đạo hàm ngẫu nhiên (SGD) được sử dụng là hàm tối ưu hóa quá trình đào tạo mô hình. Tốc độ học ban đầu cho SGD được đặt là 0,01.

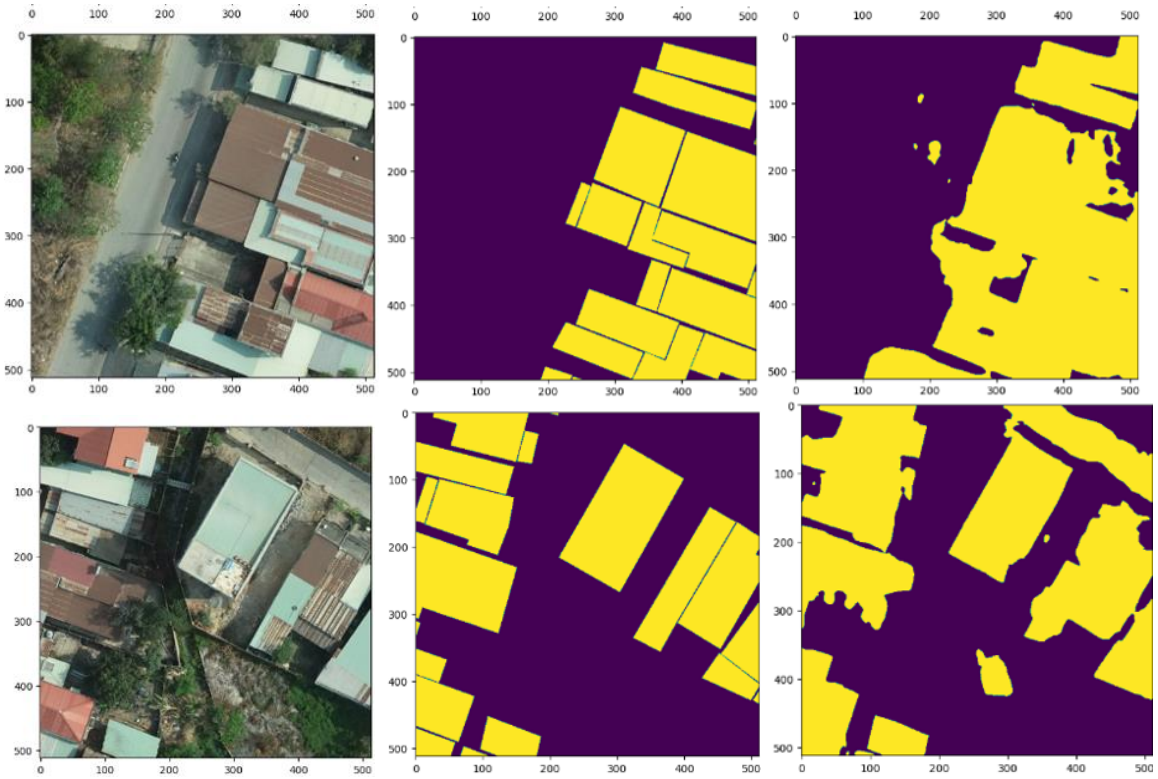
3.3. Kết quả trích xuất tòa nhà

3.3.1. Kết quả định tính

Trong nghiên cứu này, ba bộ dữ liệu sử dụng để đào tạo (training) mô hình, validation và testing được sử dụng ở khu vực có đặc điểm đối tượng cần trích xuất (lớp nhà) tương tự nhau. Mô hình được học từ 5000 mẫu ảnh có kích thước 512×512 pixel. Kết quả trích xuất đối tượng từ bộ mẫu dữ liệu được thể hiện như Hình 7 và Hình 8.



Hình 7. Kết quả trích xuất đối tượng từ bộ mẫu dữ liệu cho khu đô thị mới: Ảnh UAV (trái), nhãn đúng (giữa), nhãn mô hình dự đoán (phải).



Hình 8. Kết quả trích xuất đối tượng từ bộ mẫu dữ liệu cho khu đô thị cũ. Ảnh UAV (trái), nhãn đúng (giữa), nhãn mô hình dự đoán (phải).

Có thể thấy rằng kết quả trích xuất đối tượng trên ảnh độ phân giải cao UAV đã đạt được những yêu cầu về nhận biết đúng đối tượng. Tuy nhiên, độ chính xác tòa nhà được trích xuất ở cả khu vực đô thị cũ và mới từ mô hình U-Net đều còn hạn chế. Một số hình ảnh của hiên nhà hoặc các mái che cạnh nhà vẫn còn bị lẫn vào kết quả trích xuất tòa nhà. Một số yếu tố

khác như bóng tòa nhà, hoặc cây xung quanh nhà che khuất một phần mái nhà là nguyên nhân chính dẫn đến việc trích xuất tòa nhà còn thiếu chính xác.

3.3.2. Kết quả định lượng

Kết quả trích xuất dữ liệu tòa nhà tại khu đô thị mới và cũ đều chưa đạt độ chính xác cao. Chỉ số đánh giá mức độ phù hợp tốt nhất với bài toán phân đoạn ngữ nghĩa thường được áp dụng là IoU với mức xấp xỉ 60% (Bảng 2). Kết quả này có thể được giải thích bởi các lý do gồm: mức độ phức tạp về độ cao của tòa nhà và công trình xây dựng tại các đô thị của nước ta, ranh giới giữa tòa nhà với các đối tượng xung quanh thường không được phân biệt rõ ràng (ví dụ do cây trồng xung quanh nhà). Ngoài ra, do mật độ xây dựng cao cộng với sự đa dạng về màu sắc mái cũng như hình khối thay đổi của các tòa nhà cũng dẫn đến kết quả trích xuất còn thấp. Một nguyên nhân khác cũng không thể không kể đến đó là các vật kiến trúc khác như sân nhà, mái che cũng có hình dạng và màu sắc tương tự như tòa nhà làm cho mô hình bị nhầm lẫn ngay cả khi học và đến khi dự đoán đối tượng từ dữ liệu thực tế.

Kết quả so sánh độ chính xác tòa nhà được trích xuất trong khu đô thị mới và cũ chỉ ra rằng không có sự khác biệt đáng kể của hai khu vực nghiên cứu này. Đánh giá ở tất cả các thước đo, mức độ khác biệt chỉ từ 3 đến 7% và trung bình khoảng 4%. Kết quả trích xuất tòa nhà ở khu đô thị mới tốt hơn là do mức độ đa dạng của các tòa nhà của khu đô thị cũ rất lớn trong khi tại các khu đô thị mới các tòa nhà thường có cấu trúc tương tự nhau.

Bảng 2. Đánh giá độ chính xác trích xuất tòa nhà sử dụng mạng U-Net.

Khu vực nghiên cứu	Thước đo độ chính xác					
	Accuracy	**Precision**	**Recall**	**F1-Score**	**IoU**	**Dice**
Khu đô thị mới	0,944	0,719	0,808	0,755	0,613	0,861
Khu đô thị cũ	0,919	0,640	0,854	0,711	0,577	0,831

4. Kết luận

Mô hình mạng học sâu sử dụng bộ dữ liệu lớn (*Big data*) đang là một xu hướng được ứng dụng rộng rãi trong nhiều lĩnh vực. Sử dụng mạng học sâu tự động trích xuất đối tượng trên ảnh đã đạt được những kết quả bước đầu đầy hứa hẹn. Công nghệ này giúp giảm thiểu tối đa công tác số hóa thủ công trên ảnh UAV và đồng thời mở ra một hướng mới trong ứng dụng quản lý xây dựng quá trình đô thị hóa nhanh chóng ở nước ta hiện nay. Kết quả trích xuất tòa nhà từ dữ liệu UAV sử dụng mạng U-Net đạt độ chính xác khoảng 60%. Mô hình mạng học sâu có thể nhận biết được hầu hết mọi tòa nhà và công trình xây dựng ở trên ảnh tuy nhiên các vùng dự đoán từ mô hình chưa thật trùng khớp với các tòa trên thực tế.

Để nâng cao độ chính xác dự đoán từ mô hình có thể sử dụng các biện pháp bao gồm: (1) Nâng cao chất lượng và số lượng mẫu trong quá trình đào tạo và phát triển mô hình; (2) Sử dụng các mô hình cải tiến mới, (3) Áp dụng biện pháp học chuyển giao (transfer learning) để áp dụng có các bộ mẫu có kích thước vừa phải.

Cuối cùng, có thể khẳng định rằng việc áp dụng công nghệ trí tuệ nhân tạo trong công tác trích xuất dữ liệu nói chung và xây dựng bản đồ, cập nhật dữ liệu nền địa lý nói riêng là phù hợp. Đây là kết quả quan trọng nhằm bổ sung thêm cơ sở khoa học và có ý nghĩa thực tiễn trong thông tư 07/2021 của Bộ Tài nguyên và Môi trường về việc cho phép thành lập bản đồ địa hình tỉ lệ lớn trên cơ sở dữ liệu từ thiết bị bay không người lái UAV.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: N.H.L., P.T.D.; Xử lý số liệu: P.T.D., L.T.T.H.; Viết bản thảo bài báo: N.H.L., L.T.T.H., P.V.C., P.T.L., Đ.D.H.; Chỉnh sửa bài báo: P.T.D., L.T.T.H.

Lời cảm ơn: Nghiên cứu này được hỗ trợ bởi đề tài mã số: SPD2023.01.09.

Lời cam đoan: Tập thể tác giả cam đoan bài báo này là công trình nghiên cứu của tập thể tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Biljecki, F.; Stoter, J.; Ledoux, H.; Zlatanova, S.; Çöltekin, A.J.I.I.J.o.G.I. Applications of 3D city models: State of the art review. *ISPRS Int. J. Geo-Inf.* **2015**, *4*(4), 2842–2889.
2. Biljecki, F.; Ledoux, H.; Stoter, J.; Zhao, J.J.C. Formalisation of the level of detail in 3D city modelling. *Comput. Environ. Urban Syst.* **2014**, *48*, 1–15.
3. Akmalia, R.; Setan, H.; Majid, Z.; Suwardhi, D.; Chong, A. TLS for generating multi-lod of 3D building model. *IOP Conf. Ser. Earth Environ. Sci.* **2014**, *18*, 012064.
4. Wang, C.; Ferrando, M.; Causone, F.; Jin, X.; Zhou, X.; Shi, X.J.B. Data acquisition for urban building energy modeling: A review. *Building Environ.* **2022**, *217*, 109056.
5. Aicardi, I.; Chiabrando, F.; Grasso, N.; Lingua, A.M.; Noardo, F.; Spanò, A. UAV photogrammetry with oblique images: First analysis on data acquisition and processing. Proceeding of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B1, 2016 XXIII ISPRS Congress, 12–19 July 2016, Prague, Czech Republic, 2016, pp. 835–842.
6. Tabarro, P.; Pouliot, J.; Fortier, R.; Losier, L.M. A webgis to support GPR 3D data acquisition: A first step for the integration of underground utility networks in 3D city models. *Remote Sens. Spatial Inf. Sci.* **2017**, *XLII-4/W7*, 43–48.
7. Xiao, Y.; Zhan, Q.; Pang, Q. 3D data acquisition by terrestrial laser scanning for protection of historical buildings. 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, China, 2007, pp. 5971–5974.
8. Yang, B.J.S. Developing a mobile mapping system for 3D GIS and smart city planning. *Sustainability* **2019**, *11(13)*, 3713.
9. Singh, S.P.; Jain, K.; Mandla, V.R. Virtual 3D city modeling: Techniques and applications. *Int. Arc. Photogramm. Remote Sens. Spat. Inf. Sci. XL-2/W* **2013**, *2*, 73–91.
10. Valencia, J.; Muñoz-Nieto, A.; Rodríguez-Gonzálvez, P. Virtual modeling for cities of the future. State-of-the art and virtual modeling for cities of the future. State-of-the art an. *Int. Arc. Photogramm. Remote Sens. Spat. Inf. Sci. XL-5/W4* **2015**, 179–185.
11. Schlosser, A.D.; Szabó, G.; Bertalan, L.; Varga, Z.; Enyedi, P.; Szabó, S.J.R.S. Building extraction using orthophotos and dense point cloud derived from visual band aerial imagery based on machine learning and segmentation. *Remote Sens.* **2020**, *12(15)*, 2397.
12. Li, Y.; He, B.; Long, T.; Bai, X. Evaluation the performance of fully convolutional networks for building extraction compared with shallow models. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 2017, pp. 850–853.
13. Lari, Z.; Ebadi, H. Automated building extraction from high-resolution satellite imagery using spectral and structural information based on artificial neural networks. ISPRS Hannover Workshop, 2007, pp. 1–4.
14. Turlapaty, A.; Gokaraju, B.; Du, Q.; Younan, N.H.; Aanstoos, J.V. A hybrid approach for building extraction from spaceborne multi-angular optical imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5(1)*, 89–100.
15. Chen, R.; Li, X.; Li, J. Object-Based Features for House Detection from RGB High-Resolution Images. *Remote Sens.* **2018**, *10*, 451.
16. Zhu, Y.; Huang, B.; Gao, J.; Huang, E.; Chen, H. Adaptive polygon generation algorithm for automatic building extraction. *IEEE Int. Geosci. Remote Sens. Symp. Proc.* **2021**, *60*, 1–14.

17. Yin, J.; Wu, F.; Qiu, Y.; Li, A.; Liu, C.; Gong, X.J.R.S. A multiscale and multitask deep learning framework for automatic building extraction. *Remote Sens.* **2022**, *14(19)*, 4744.
18. Wen, Q.; Jiang, K.; Wang, W.; Liu, Q.; Guo, Q.; Li, L.; Wang, P. Automatic building extraction from google earth images under complex backgrounds based on deep instance segmentation network. *Sensors* **2019**, *19(2)*, 333.
19. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10(1)*, 144.
20. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. Proceeding of the Conference: The Handbook of Brain Theory and Neural Network, 1995, pp. 1–14.
21. Hu, Q.; Zhen, L.; Mao, Y.; Zhou, X.; Zhou, G. Automated building extraction using satellite remote sensing imagery. *Autom. Constr.* **2021**, *123*, 103509.
22. Deng, X.; Liang, Y.; Li, X.; Xu, W. Recognition and spatial distribution of rural buildings in Vietnam. *Land* **2023**, *12(12)*, 2142.
23. Nguyen, A.; Luu, H.; Phan, A.; Bui, H.; Nguyen, T. Cau Giay: A dataset for very dense building extraction from google earth imagery. 2019 6th NAFOSTED Conference on Information and Computer Science (NICS), Hanoi, Vietnam, 2019, pp. 352–356.
24. Li, Z.; Xin, Q.; Sun, Y.; Cao, M. A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery. *Remote Sens.* **2021**, *13*, 3630.
25. Sun, X. Deep learning-based building extraction using aerial images and digital surface models, Geoinformation Science and Earth Observation MSc, University of Twente, 2021.
26. Ndung'u, R.N. Data preparation for machine learning modelling. *Int. J. Comput. Appl. Technol. Res.* **2022**, *11(06)*, 231–235.
27. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. Proceeding of the International Conference on Engineering and Technology (ICET), 2017.
28. Terven, J.; Cordova-Esparza, D.M.; Ramirez-Pedraza, A.; Chavez-Urbiola, E.A.; Romero-Gonzalez, J. Loss functions and metrics in deep learning. A review. *Computer Sci*. **2023**, 1–85.
29. Boursalie, O.; Samavi, R.; Doyle, T.E. Evaluation metrics for deep learning imputation models. International Workshop on Health Intelligence, Springer, 2021.
30. Lê, V.C.; Lê, T.T.H.; Cao, X.C. Nghiên cứu lựa chọn vị trí cất cánh cho thiết bị bay không người lái tích hợp GNSS động phục vụ đo vẽ thành lập bản đồ địa hình tỷ lệ lớn cho các mỏ lộ thiên. *Tạp chí Khoa học kỹ thuật Mỏ - Địa chất* **2020**, *61(5)*, 54–63.
31. Long, N.Q.; Cường, C.X. Ứng dụng máy bay không người lái (UAV) để xây dựng mô hình số bề mặt và bản đồ mỏ lộ thiên khai thác vật liệu xây dựng. *Tạp chí Khoa học kỹ thuật Mỏ - Địa chất* **2020**, *61(1)*, 21–29.
32. Sỹ, M.V.; Quý, B.N.; Hiệp, P.V.; Quý, L.Đ. Nghiên cứu sử dụng dữ liệu ảnh máy bay không người lái (UAV) trong thành lập bản đồ địa hình tỷ lệ lớn. *Tạp chí Khoa học Đo đạc và Bản đồ* **2017**, *33*, 49–57.
33. Cường, N.S.; Trường, T.X.; Hạnh, T.H.; Vũ, Đ.N. Nâng cao chất lượng xây dựng mô hình 3D bằng kết hợp công nghệ bay chụp UAV và quét laser mặt đất. *Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất* **2019**, *60(4)*, 31–40.
34. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N.; Hornegger, J.; Wells, W.; Frangi, A.

(eds) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, Springer, Cham. 2015, 9351, 234–241.
35. Naidu, G.; Zuva, T.; Sibanda, E.M. A Review of evaluation metrics in machine learning algorithms. In: Silhavy, R., Silhavy, P. (eds) Artificial Intelligence Application in Networks and Systems. CSOC 2023. Lecture Notes in Networks and Systems, Springer, Cham. 2023, 724, pp. 15–25.
36. Rainio, O.; Teuho, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep*. **2024**, *14(1)*, 6086.
37. Jadon, S. A survey of loss functions for semantic segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Via del Mar, Chile, 2020, pp. 1–7.
38. Punn, N.S.; Agarwal, S. Inception U-Net architecture for semantic segmentation to identify nuclei in microscopy cell images. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), **2020**, *16(1)*, 1–15.

Building extraction from UAV images using deep learning for urban management and establishment of 3D LoD1

Nguyen Huu Long[1], Pham Van Chung[2], Pham Thi Lan[2], Nguyen Van Trung[2,4], Le Thi Thu Ha[2,4], Dang Dieu Hue[3], Pham Trung Dung[2]*

[1] Dong Thap University; nhlong@dthu.edu.vn
[2] Hanoi University of Mining and Geology; phamvanchung@humg.edu.vn; nguyenvantrung@humg.edu.vn; lethithuha@humg.edu.vn; phamtrungdung@humg.edu.vn
[3] Hanoi University of Civil Engineering; huedd@huce.edu.vn
[4] Geomatics in Earth Sciences Research Group, Hanoi University of Mining and Geology; nguyenvantrung@humg.edu.vn; lethithuha@humg.edu.vn

Abstract: Automatic building extraction from UAV (Unmanned Aerial Vehicle) images involves identifying and vectorizing buildings directly from image data. In urban management and development, it is essential to store buildings and architectural structures in a geo-database and visualize them on a computer. This paper introduces a novel method that utilizes deep learning technology to automatically extract the roofs of buildings and construction projects from UAV images. The extracted roof data is combined with height and other relevant information about the structures to establish and update 3D Level of Detail (LoD1) models. We tested the building extraction using the U-Net deep learning model in both new and older urban areas, achieving an accuracy of 60%. The findings of this study can be applied to construction management in cities experiencing rapid urbanization in our country today.

Keywords: Automatically building extraction; Deep learning; 3D model LoD1; Urban management; UAV.