衛星画像と異種地理情報の統合に基づくU-Netによる樹冠高推定

内藤　洋輝†　　桑田　若菜†　　大島　裕明†

† 兵庫県立大学 情報科学研究科　〒651–2197 神戸市西区学園西町 8–2–1
E-mail: †ad25a046@guh.u-hyogo.ac.jp, †af25x004@guh.u-hyogo.ac.jp, †ohshima@ai.u-hyogo.ac.jp

あらまし　本研究では，衛星画像を入力として，各画素に対応する樹冠高を推定する手法を提案する．樹冠高は，森林の分布や状態を把握する上で重要な指標であり，これまでは衛星画像を用いて樹冠高を推定する研究が行われてきた．しかし，樹冠高は標高や気候，緯度経度といった地理的要因の影響を強く受けることが知られており，衛星画像のみを入力とする手法では，これらの地理的要因を考慮できない．そこで本研究では，衛星画像に加えて地理的要因を追加特徴量として統合する樹冠高推定手法を提案する．さらに，地理的要因をどのように機械学習モデルへ統合すれば効果的であるかを検証し，樹冠高推定における地理情報統合の有効性を明らかにする．

キーワード　衛星画像，社会インフラデータ，U-Net

1　はじめに

森林は地球の陸地の約 3 割を占めており，二酸化炭素の吸収や気温，水循環の調整など，気候変動の安定化において重要な役割を果たしている．その中でも，樹冠高は森林の質や機能を評価する上で注目される指標の一つである．樹冠とは樹木のうち葉と枝の集まった部分，樹冠高とは地面から樹冠の最も高い部分までの垂直距離と定義される．樹冠高は場所や環境によって大きく異なる．近年の研究により，樹冠高は標高，緯度，気候帯などの地理的要因から強い制約を受けるということが明らかになっている [1], [7], [11].

樹冠高は環境面だけではなく社会インフラにも影響を与える．例えば通信分野では，樹冠の密度や高さによって携帯電話の電波強度が変化し，特に高樹が密集する地域では電波の減衰や遮断を引き起こす可能性がある．このように, 樹冠高は自然環境の理解や社会インフラの最適化を考えるうえで重要な指標であり，樹冠高の推定は，地球環境の理解や電波遮断地域の特定といった社会的価値に繋がる．

樹冠高推定の既存研究として, 視差情報を利用した手法 [3], [17] がある．この手法では，ドローンによって撮影された複数の角度の航空写真を利用し，それらの視差を用いて地表面と樹冠表面を含む 3 次元復元を行う．この復元結果の 3 次元点群の差分として，樹冠高を算出している．そのため，樹冠高を直接推定しているわけではなく，地表面の推定誤差や植生以外の構造物の影響が樹冠高の推定性能に影響を与えてしまうといった問題がある．また，この手法は複数視点の画像を前提としているため，ドローンの運用コストや天候条件，風による影響の制約を受けやすいことが指摘されている [14].

近年では，リモートセンシングや画像認識モデルの発展により，LiDAR や衛星画像を用いて樹冠高を推定する研究が行われている．しかし，既存の高解像度な樹冠高データは，過去の観測結果をもとに作成されたものしか存在しない．一方で，衛星画像は現在も継続的に取得されており，最新の地表状態を広域かつ高頻度に観測できる．樹冠高は成長や伐採，災害などにより時間とともに変化するため，過去に取得された樹冠高データのみでは現時点における森林構造を正確に把握することは困難である．特に，通信インフラ設計や森林管理といった実社会の応用においては，最新の樹冠高を推定することへの需要が高まっている．

LiDAR は高い性能を期待できるが，高額なコストがかかるため，観測範囲や頻度が限られてしまう．一方で，衛星画像は単一画像で広範囲をモニタリングでき，データの利用も容易である．そのため，本研究では衛星画像を利用する．しかし，衛星画像のみを入力とした場合，標高や緯度，気候帯といった地理的要因を考慮できないという制限があり，これらの地理情報を考慮することで樹冠高の推定性能が向上する可能性がある．

そこで本研究では，衛星画像だけでなく，地理情報を追加特徴量として組み込む機械学習モデルを構築し，樹冠高推定の性能向上を目指す．

本研究では以下の 3 つのリサーチクエスチョン（RQ）を設定し，それぞれに対する実験を通じて，地理情報に基づく樹冠高の推定の可能性を明らかにする．

RQ1　衛星画像から樹冠高をどの程度推定することができるか？

RQ2　標高，緯度，気候などの地理情報を考慮することで，樹冠高の推定性能は向上するのか？

RQ3　樹冠高の推定において，地理情報を効果的に考慮するには，どのような方法が最適か？

2　関連研究

2.1　衛星画像からの樹冠高の推定

Jamie ら [16] は，衛星画像を事前学習した Vision Transformer（ViT）[4] を特徴抽出器として用い，その出力を畳み込み型デコーダに入力することで地上解像度 1m での樹冠高推定を実現している．この研究では，カリフォルニア州とサンパウロ州が対象地域として選定されている．特にカリフォルニア州

では 2018 年から 2020 年にかけて撮影された高解像度衛星画像を用いて州全域にわたる樹冠高マップを生成している．また，Wagner ら [19] は，カリフォルニア州全域を対象に，U-Net [13] を用いて衛星画像から画素ごとの樹冠高を推定している．

これらの研究では，画素全体を対象とした平均絶対誤差（MAE）を用いて評価が行われており，樹木が存在しない画素も含めた形で誤差が算出されている．そのため，樹木が存在する領域における樹冠高推定性能のみに着目した詳細な評価は行われていない．また，これらの手法は衛星画像の画像特徴のみに基づく推定であり，標高や気候条件といった樹冠高に影響を及ぼし得る地理的要因を考慮していない．

2.2 地理情報が樹冠高に及ぼす影響

樹冠高は，単に樹種の特性だけで決まるわけではなく，標高，緯度経度，気候帯などといった地理的要因から強い制約を受けることが知られている．近年では，これらの地理的要因と樹冠高の関係を広域的に分析する研究が数多く報告されている．

Gelabert ら [7] と Ameztegui ら [2] は，ヨーロッパの山岳地域を対象として標高と最大樹冠高の関係を分析した．彼らの研究では，ある標高までは樹冠高の変化は小さいものの，特定の標高を境に樹冠高が低下するしきい値が存在することを発見した．これは，標高は一定の値を境に，樹木がそれ以上高く成長できなくなる境界として機能することを示している．

Rahimi ら [11] は，地球規模での樹冠高と地表面温度の関係に着目し，ある一定の樹冠高から地表面温度が低下する非線形関係があることを発見した．また，緯度帯によってその強さや傾向が異なり，特に熱帯地域が最も強い関係があると報告している．

Adrah ら [1] や Fricker ら [6] は，それぞれマレーシア全域とカリフォルニア州を対象に，勾配ブースティングモデル [9] を用いて標高や気候情報から空間解像度ごとの最大樹冠高を推定した．その結果，年平均降水量が樹冠高の推定において最も影響力の大きい決定因子であることを判明し，加えて年平均気温や標高も一定の影響を与えることを示した．このように，樹冠高は気候条件や標高等の地理情報から強い制約を受けることが明らかになっている．

2.3 U-Net

Ronneberger らが提案した U-Net [13] は，生物医学画像分野におけるセグメンテーションを目的として提案されたモデルである．U-Net は画像の全体的な特徴を抽出するためのエンコーダ部と，特徴マップの解像度を段階的に上げていくデコーダ部から構成される．エンコーダ側で抽出された特徴マップをスキップ接続を介してデコーダ側に組み合わせることで，より正確な位置特定を可能としている．このような構造により，U-Net はセグメンテーションのような画素単位で値やクラスを推定するタスクで多く活用されている．衛星画像を入力とした樹冠高推定においても，画素ごとに値を推定できる点で有力なベースラインモデルとなり得る．本研究では，衛星画像から各ピクセルの樹冠高を推定するためのベースラインモデルとして U-Net を用いる．

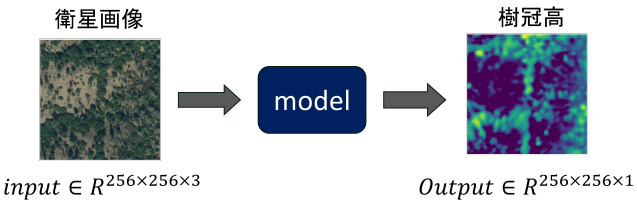

図 1: 入出力例

Stable Diffusion [12] は，U-Net に Cross-Attention 機構 [18] を導入し，画像特徴とテキスト情報を柔軟に統合することを実現している．この手法では，画像特徴を Query，テキストの条件入力を Key および Value とする Cross-Attention を適用する．これにより，画像の各画素の特徴が，条件入力のどの要素に注目すべきかを動的に学習でき，条件情報を画像特徴に効果的に統合することができる．出力される Cross Attention Weight は各画素に対する条件入力の注目度を表し，拡散モデルにおける視覚言語現象の解析 [10], [15] やセグメンテーション技術 [20] などに活用されている．

Cross-Attention はエンコーダ側の低解像度の画像特徴抽出からデコーダ側の復元過程まで適応される．そのため，一貫して条件情報を反映した画像特徴の抽出が可能となる．

2.4 本研究の位置づけ

以上を踏まえ，本研究では衛星画像に加えて標高や気候といった地理情報を追加特徴量として統合する機械学習モデルを構築し，樹冠高の推定性能の向上を目指す．具体的には，衛星画像から抽出される視覚的特徴と，緯度経度，標高，気温，降水量などの数値的な地理情報を統合することで，画像情報のみでは捉えきれない地理的要因を考慮した推定を実現する．また，本研究では地理情報をどのように機械学習モデルに組み込むことが予測性能の向上につながるのかを調査する．

3 問題定義とデータセット

3.1 問 題 定 義

本研究では，衛星画像を入力として，各画素に対応する樹冠高を推定する機械学習モデルの構築を目的とする．本問題は，以下のように定義される．

入力 RGB 衛星画像 $X \in \mathbb{R}^{256\times256\times3}$

出力 各画素における樹冠高 $Y \in \mathbb{R}^{256\times256\times1}$

本タスクは，衛星画像に含まれる視覚情報を手がかりに，対応する各画素の樹冠高を推定する回帰問題である．図 1 にこの問題の入出力例を示す．

3.2 衛星画像と樹冠高データ

本研究では，入力する衛星画像として，National Agriculture Imagery Program（NAIP）を使用する．NAIP はアメリカ合衆国アメリカ合衆国本土を対象に取得された地上解像度 1m の衛星画像である．これは，1 画素が地上約 1m 四方の領域に対応することを意味する．NAIP は 2003 年から現在まで撮影さ

れており，2009 年からは州単位で 3 年以下の間隔で撮影されている．NAIP は州単位で DOQQ と呼ばれる単位に分割されており，各 DOQQ には撮影日を含むメタデータが付与されている．本研究では，この撮影日を後述する樹冠高データとの時間整合性を確保するために利用した．

一方，正解データとなる樹冠高には，Meta が公開している Canopy Height Map（CHM）を使用した．CHM は地上解像度 1m で提供される樹冠高マップであり，2009 年から 2020 年までの全世界の樹冠高を提供している．その中でもデータの 80%は 2018 年から 2020 年に取得されており，近年の森林構造を反映している．

CHM はタイルと呼ばれる単位に分割されており，各タイルは $65,536 \times 65,536$ ピクセルから構成される．各タイルはさらに細分化された領域ごとに分割されており，各領域には樹冠高の取得日を表すメタデータが付与されている．このため，CHM は同一タイル内であっても観測時期が一様ではない点に注意が必要である．

本研究では，CHM の主要な生成期間を考慮し，2018 年から 2020 年を対象期間とした．また，NAIP の撮影日と CHM の観測日の差が小さいペアのみを選択し，時間的に整合性の高いペアデータセットを構築した．

3.3 ペアデータセットの作成

本研究では，アメリカ合衆国カリフォルニア州を対象地域とし，NAIP と CHM を同一地点かつ近接した取得日の組み合わせでペアリングし，学習に適した大規模データセットを構築した．カリフォルニア州は気候条件が多様であり，山岳地帯も有することから，衛星画像と地理情報を用いた樹冠高推定に有効性を検証する上で適した地域である．NAIP と CHM の取得日の差が大きい場合，樹木の成長や伐採などの影響により学習データとして不整合が生じる可能性がある．そこで本研究では，NAIP の撮影日と CHM の観測時期の差が 30 日以内となる組み合わせのみをペアデータセットとして採用した．ペアデータセットの作成手順を以下に示す．

3.3.1 CHM タイルと NAIP DOQQ の取得

まず，カリフォルニア州境界ポリゴンを用いて，州内に含まれる CHM タイルを抽出した．具体的には，CHM タイルの空間範囲とカリフォルニア州境界ポリゴンとの空間的な重なり判定を行い，州内に含まれる CHM タイルを対象とした．その結果，カリフォルニア州内では 143 個の CHM タイルが対象となった．

次に，NAIP Image Dates [1] から，2018 年および 2020 年のカリフォルニア州の NAIP DOQQ を取得した．NAIP Image Dates は各 DOQQ がカバーする空間範囲と撮影日を対応付けたメタデータを提供する公開サイトである．本研究ではこの撮影日情報を CHM の観測日との時間的整合性を確保するために利用した．

1：https://naip-image-dates-usdaonline.hub.arcgis.com/

3.3.2 日付差 30 日以内のフィルタリング

まず，CHM については，タイル内で細分化された各領域に付与された観測日メタデータを取得した．CHM は同一タイル内であっても，領域ごとに観測日が異なるため，各領域を独立した単位として扱った．次に，2018 年および 2020 年に撮影された NAIP DOQQ のメタデータを読み込み，各 DOQQ に付与された衛星画像の撮影日メタデータを取得した．その後，CHM の各細分領域と空間的に重なる NAIP DOQQ を抽出し，両者の取得日の差を算出した．

最後に，取得日の差が 30 日以内である組み合わせのみを残し，該当する CHM 範囲と NAIP 範囲の共通領域をポリゴンとして抽出した．このポリゴンを後続の画像パッチ生成処理における有効領域として利用した．

3.3.3 CHM 画像パッチの生成

前節で抽出した CHM と NAIP の共通領域ポリゴンを用いて学習用の画像パッチを作成した．まず，各 CHM タイル内で共通領域に完全に含まれる画素領域を特定した．その後，共通領域内に完全に含まれる領域のみを対象として，地上解像度 1m に対応する 256×256 ピクセルの CHM 画像パッチを切り出した．このとき，共通領域から一部でもはみ出すパッチや，欠損値を含むパッチは除外した．この処理により，カリフォルニア州全体で合計 447,946 件の CHM 画像パッチを作成した．また，各パッチについて，中心の緯度経度を算出し，CHM のメタデータと空間的に対応付けることで，樹冠高の観測日を取得した．

3.3.4 NAIP 画像パッチの生成

前節で作成した CHM 画像パッチに対して，対応する NAIP 画像を取得した．各 CHM パッチは，中心の緯度経度、観測日をデータとして保存しており，これらの情報を用いて空間的・時間的に対応する NAIP 画像を選択した．具体的には，各 CHM パッチの中心の緯度経度を利用して，その緯度経度から地上解像度 1m に対応する 256×256 ピクセルの NAIP 画像を Google Earth Engine の API を用いて取得した．NAIP 画像は，2018 年から 2020 年の期間に撮影された DOQQ データを対象とし，RGB の 3 バンドを使用した．複数の NAIP 画像が同一領域を覆う場合には，CHM パッチの観測日との差が最小となる NAIP 画像を選択した．このとき，NAIP の撮影日と CHM の観測日の差が 30 日以内である場合のみを有効な対応関係として採用した．条件を満たさない場合や，対応する NAIP 画像が存在しない場合には，当該パッチを除外した．その結果，CHM 画像パッチと時間的・空間的に対応付けられた NAIP 画像パッチは合計 396,556 件となった．

以上の処理により，CHM および NAIP の大規模な画像ペアデータセットを構築した．しかし，計算資源および学習時間を考慮し，本研究では，これらのペアデータセットの中から 10,000 件をランダムサンプリングし，実験に使用した．サンプリングした 10,000 件のペアデータセットの例を図 2 に示す．図 2 に示す CHM 画像パッチのカラーマップは，可視性を高めるために，各 CHM 画像パッチごとに最小値と最大値に基づいて表示している．そのため，画像パッチ内の色は同一パッチ内

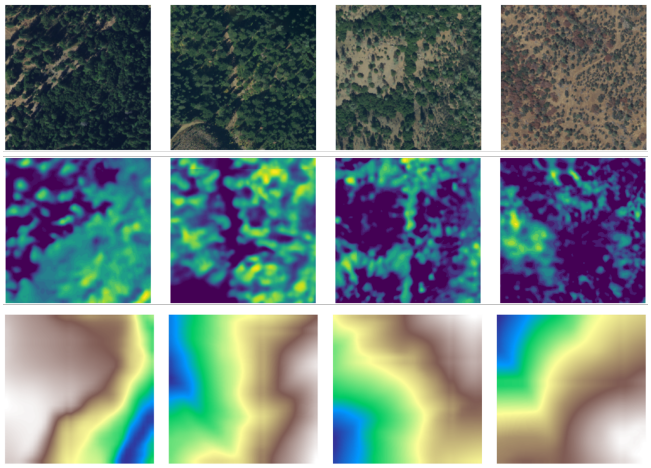

図 2: ペアデータセットの例
上：NAIP 画像 中：CHM 画像 下：標高画像

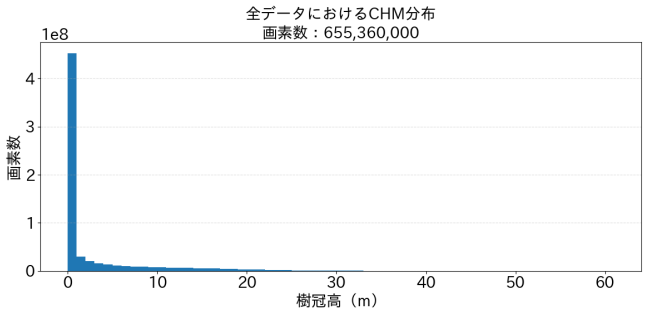

図 3: 10,000 件の CHM 画像パッチ内の
全画素の樹冠高分布

表 1: CHM データセットにおける樹冠高値の画素分布

	全画素数	値 >0 の画素数	値 $=0$ の画素数
全体	655,360,000	202,598,679	452,761,321
訓練	524,288,000	162,024,862	362,263,138
検証	65,536,000	19,416,832	46,119,168
テスト	65,536,000	21,156,985	44,379,015

における相対的な高低のみを表している．

CHM 画像パッチ内の全画素の樹冠高分布を図 3 に示す．図 3 より，樹冠高が 0 である画素が大部分を占め，樹木が存在しない領域が広く含まれていることが分かる．

抽出した 10,000 件のデータについては，ランダムに学習用データ 8,000 件，検証用データ 1,000 件，テスト用データ 1,000 件に分割した．分割した CHM データセットに含まれる，値が 0 である画素数および 0 より大きい画素数の内訳を表 1 に示す．

3.4 地理情報データセット

樹冠高は，樹木が生育する地形条件や長期的な環境条件の影響を強く受ける．例えば，標高は最大樹冠高を決定し，気温や降水量，日射量といった気候条件は，樹木の成長や樹冠構造の形成に直接的な影響を与える．

そこで本研究では，衛星画像に加えて，地形条件および環境条件を表現する補助的な地理情報として，標高データと気象データを導入した．標高データは空間的な分布を持つ画像情報として扱い，気象データは各地点に対応する数値情報として扱う．これにより，樹冠高推定において，画像情報のみでは捉えにくい地理的・環境的要因を考慮することを可能とした．

3.4.1 標高データ

標高データには，USGS が公開している 3D Elevation Program（3DEP）の 10m 解像度データを使用した．このデータは Google Earth Engine 上で利用可能であり，広域な標高情報を取得できる．

3.2 節で作成した 10,000 件の各 CHM 画像パッチの中心の緯度経度を基準として，一辺約 256m の正方形領域を定義し，10m 解像度で標高をサンプリングした．取得された標高データは，元の空間解像度では約 26×26 ピクセルとなるため，CHM および NAIP 画像と空間サイズを統一する目的で，バイリニア補間を用いて 256×256 ピクセルへアップサンプリングした．バイリニア補間は，ある画素の値を周囲 4 つの画素の値を用いて距離に応じて平均を取ることで値を補間する．これにより，すべての入力データが同一の画素数を持つように整形した．図 2 に NAIP 画像パッチおよび CHM 画像パッチと同一の中心座標から生成した標高画像パッチを示す．標高画像パッチも CHM 画像パッチ同様，可視性を高めるために各パッチごとに最小値と最大値に基づいてカラーマップを表示している．

本研究では前処理として各画素の標高の値を 1,000 で割る処理を施した．これにより，標高値のスケールを適度な範囲に収め，衛星画像と組み合わせた際のスケール差を緩和させた．

3.4.2 気象データ

気象データには，NASA が提供する Daymet V4 を使用した．Daymet V4 は，北米を対象とした高解像度の地上気象データセットであり，日次の気象変数を 1km 解像度で提供している．Daymet V4 は本研究で使用する NAIP および CHM とは空間解像度が異なるが，気温や降水量といった気候条件が 1m 単位で急激に変化することは考えにくい．そのため，本研究では Daymet V4 を元の 1km 解像度のまま利用する．

本研究では，Daymet V4 に含まれる以下の 7 種類の気象変数を使用した．これらの変数は，樹木の生育環境を総合的に表現できる指標である．

- **最大気温**
- **最低気温**
- **降水量**
- **短波放射量**
- **水蒸気圧**
- **積雪水量**
- **日照時間**

Daymet V4 は日次の気象データであるが，本研究では，単一日の気象条件ではなく，長期的な環境条件が樹冠高に与える影響を考慮するため，気象データを疑似的な気候データとして扱った．具体的には，各 CHM 画像パッチの観測日を基準とし，その日から過去 1 年間の期間について，各気象変数の平均値を算出した．この 1 年平均化により，短期的な天候変動ではなく，長期的な気候情報を抽出した．

3.5 数値情報の前処理

本研究では各画像パッチの中心の緯度経度および Daymet V4 の各気候情報の 9 種類を数値情報として扱う．これらの数値情報について，スケールを均一化するために以下の前処理を施した．まず，緯度経度については，それぞれ 90 および 180 で除算する処理を施し，位置関係を保持したまま値域を下げた．Daymet V4 の各気候情報については，日照時間以外の変数については Z スコア標準化を行い，平均 0，分散 1 となるように変換した．日照時間については 10,000 で除算する処理を施した．これによりモデルの学習の安定化だけでなく，単位や値域の異なる気候情報を同等に扱えるようにした．各画像パッチには，対応する位置および気候条件を表すこれらの 9 次元の数値ベクトルが紐づけられる．

4 衛星画像からの樹冠高の推定

本節では，樹冠高推定のための機械学習モデルの概要と，地理情報を追加特徴量として入力する手法について説明する．まずは衛星画像のみを入力する場合をベースライン手法とする．次に，標高や気候といった地理情報を追加特徴量として組み込み，衛星画像のみでは捉えにくい地理的要因を考慮する手法を提案する．また，本研究では地理情報を画像特徴に統合するために複数の統合手法を検討した．

4.1 衛星画像からの樹冠高の推定

本研究の目的は， 3.1 節で示した通り，入力となる衛星画像から画素ごとに樹冠高を推定することである．本研究では，入力と同解像度の予測値を出力できる U-Net（`U-Net（3ch）`）を用いた．本研究で扱った U-Net の構造を図 4 に示す．U-Net はエンコーダとデコーダから構成され，エンコーダでは畳み込みと最大プーリングにより解像度を段階的に低下させながら特徴抽出を行う．畳み込みブロックは 3×3 の畳み込み層，Batch Normalization，ReLU からなる一連の処理を 2 回繰り返す構成となっている．ボトルネック層では，最も抽象度の高い特徴を抽出する．デコーダでは，転置畳み込みにより特徴マップを段階的にアップサンプリングし，出力解像度を段階的に上げる．このとき，エンコーダ側の同解像度の特徴マップをスキップ接続により結合する．具体的には，各段でアップサンプリング後の特徴と対応するエンコーダ特徴をチャンネル方向に連結し，畳み込みブロックにより統合する．これにより，ボトルネック層で得られる大域的文脈と，エンコーダの浅い層で保持される位置情報や輪郭情報を同時に活用することができる．本研究では，エンコーダの各層で 64，128，256，512，1024 と段階的に特徴次元を増加させた．また，デコーダの各層では転置畳み込み [5] により画像サイズを拡大させている．最終層では出力チャンネル数を 1 に変換し，各画素の樹冠高を表す予測値を出力する．損失関数には平均二乗誤差（MSE）を用いた．

4.2 地理情報と衛星画像からの樹冠高の推定

前節では，U-Net を用いて衛星画像から樹冠高を推定するベースラインモデルを示した．しかし，樹冠高は標高や気候条件といった地理的要因から強い制約を受けることが知られており，衛星画像の視覚情報だけではこれらの要因を十分に捉えられない可能性がある．そこで本節では，衛星画像の特徴と数値特徴量として与えられる地理情報を統合的に扱う枠組みを導入する．しかし，U-Net の画像特徴に対して地理情報をどのように統合すれば最も効果的に樹冠高推定性能を向上させられるかについては，これまで十分に検討されていない．そこで本研究では，地理情報を画像特徴に統合する複数の手法を提案し，それらの性能を比較することで本タスクに適した統合方法について検討する．

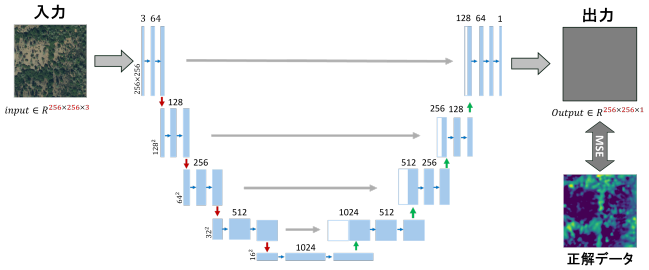

図 4: U-Net のアーキテクチャ

本節で言及する各手法に共通する操作として，衛星画像の 3 チャンネルに対して，対応する標高情報を 1 チャンネルとして空間方向に結合し，4 チャンネル入力とした．これにより，各画像パッチの各画素に対応する標高情報を視覚情報と同時にモデルへ入力することが可能となる．

4.2.1 ボトルネック層における地理情報の結合

各画像パッチに対応する地理情報は， 3.5 節で示した緯度・経度および気候変数からなる 9 次元の数値ベクトルとして与えられる．本節では，この 9 次元の地理情報ベクトルを U-Net のボトルネック層で統合する手法（`Bottleneck Concat`）を提案する．ボトルネック層の特徴は，広域的な地形構造を反映した抽象的な表現である．さらに，ボトルネック層では空間解像度が最も低く，特徴次元が高いため，1×1 畳み込みによるチャンネル方向の特徴統合を効率的に行うことが可能である．

具体的には，U-Net のエンコーダから得られるボトルネック特徴 $\boldsymbol{b} \in \mathbb{R}^{B \times 1024 \times 16 \times 16}$ に対し，地理情報ベクトルを空間方向に複製して $\boldsymbol{g} \in \mathbb{R}^{B \times 9 \times 16 \times 16}$ を生成する．これをチャンネル方向に連結することで $\boldsymbol{b}' \in \mathbb{R}^{B \times 1033 \times 16 \times 16}$ を得る．これにより，各空間位置の特徴が同一の地理条件に条件付けられ，広域的な環境要因を考慮した表現学習が可能となる．$\boldsymbol{b}'$ に対して 1×1 畳み込みを施し，元の特徴次元 $\boldsymbol{b}'' \in \mathbb{R}^{B \times 1024 \times 16 \times 16}$ に戻した上でデコーダへ入力する．この一連の操作を通じて画像特徴と地理情報の統合を行う．

4.2.2 Cross-Attention を用いた地理情報の統合

本節では，条件付き生成モデルの分野で有効性が示されている Cross-Attention 機構を応用する．Cross-Attention 機構は 2.3 節で示した通り，画像特徴を Query，条件入力を Key および Value として用いることで，画像の各画素の特徴が条件入力のどの要素に注目すべきかを動的に学習することができる．以下に Cross-Attention を用いた地理情報統合の処理手順

を示す．

まず，各画像パッチに対応する地理情報は 9 次元の数値ベクトル $\mathbf{c} \in \mathbb{R}^9$ として与えられる．各次元を 1 つのトークンとみなし，$\mathbf{c}$ を 9 トークンの系列へ変換する．具体的には，各成分 c_i をスカラーとして取り出し，線形変換により Attention の埋め込み次元 d_a へ射影することで，

$$\mathbf{t}_i = f_\theta(c_i) \in \mathbb{R}^{d_a} \quad (i = 1, \ldots, 9)$$

を得る．ここで f_θ は学習可能な全結合層である．このとき，地理情報トークン列 $T = [\mathbf{t}_1, \ldots, \mathbf{t}_9] \in \mathbb{R}^{9 \times d_a}$ を Key および Value として用いる．

次に, U-Net のある層における入力特徴マップ $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ に対して，畳み込みにより特徴を抽出し，$\mathbf{F}' \in \mathbb{R}^{C' \times H \times W}$ を得る．これを空間方向に平坦化して，$N = H \times W$ 個の位置に対応する特徴列 $\Phi \in \mathbb{R}^{N \times C'}$ へ変換する．N は画素数を表す．Φ を線形変換して Query $Q \in \mathbb{R}^{N \times d_a}$ を構成する．地理情報トークン列 T に対しても，線形変換して Key $K \in \mathbb{R}^{9 \times d_a}$，Value $V \in \mathbb{R}^{9 \times d_a}$ を構成する．これにより，画像特徴と地理情報は同一の特徴空間上に射影され，互いの類似度を計算できる表現へと変換される．

Cross-Attention は Multi-Head Attention により

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_a}}\right) V$$

として計算される．

ここで，$QK^\top \in \mathbb{R}^{N \times 9}$ は，各画素位置に対応する画像特徴と，各地理情報トークンとの類似度を表す行列である．この行列の (i, j) 成分は，i 番目の画素位置が j 番目の地理情報成分に対してどの程度注目しているかを示す重みを表す [8]．この重みは softmax 関数により，各画素における重みの総和が 1 となるように調整される．この正規化された重み行列を Attention map $\mathbf{A} \in \mathbb{R}^{N \times 9}$ とする．$\mathbf{A}$ は Value V と行列積を取ることで，各画素位置において地理情報を注意重みに基づいて重み付けした特徴を算出する．すなわち，Cross-Attention の出力は，各画素位置に対して地理情報 9 次元を注意重みに基づいて統合した条件付き特徴表現となる．得られた出力は線形変換により元のチャンネル次元 C' へ戻し，$\mathbf{F}$ に残差接続で加算した後，畳み込みにより統合特徴を得る．これにより，各画素位置の特徴が，地理情報 9 次元のうちどの要素に注目すべきかを学習しながら衛星画像特徴へ地理情報を注入できる．

本手法では, 畳み込み層, 線形変換層 f_θ, Query, Key, Value の射影行列，Attention 出力の射影層を含むすべての学習可能パラメータを学習対象とし，予測された樹冠高と正解値との平均二乗誤差（MSE）を損失関数として用いて学習を行った．

本研究では，Cross-Attention を導入する位置と数が異なる 2 種類の構成を検討した．1 つ目はボトルネック層のみに Cross-Attention を導入する構成（`Cross-Attn（Bottleneck）`）であり，抽象度の高い特徴に対して地理情報を統合する．2 つ目はエンコーダ・ボトルネック・デコーダのすべての層に Cross-Attention を導入する構成（`Cross-Attn（ALL Layer）`）であり，複数の解像度において地理情報との関係を学習できるようにした．

5 実　　験

本節では，1 節で述べた 3 つのリサーチクエスチョン（RQ）に取り組むために，衛星画像のみを入力としたベースライン手法と地理情報を追加特徴量として統合した手法との比較を行った．まず，実験設定および評価方法について述べる．その後，各手法の結果を比較し，地理情報の導入による性能向上の有無と，統合手法の違いが推定性能に与える影響を検証する．

5.1 実 験 設 定

本節では，樹冠高の推定の学習における実験設定について述べる． 4 節で示した 4 つの手法の性能を公平に比較するためにシード値を固定し，データ分割，モデルの初期化，ミニバッチの生成順序などに起因するランダム性の影響を可能な限り排除した．これにより，モデル構造や入力情報の違いによる性能差を適切に評価できるようにした．

本研究ではいずれの手法においてもシード値を 42 とし，学習パラメータとして最大エポック数を 100，Patience を 10，学習率を 1.0×10^{-5} と設定した．また，バッチサイズについては，`U-Net（3ch）`，`Bottleneck Concat`，および `Cross-Attn（Bottleneck）` の各手法では 50 とし, `Cross-Attn（ALL Layer）` では 10 とした．Cross-Attention における埋め込み次元 d_a は 1024，ヘッドの数は 4 とした．

5.2 評 価 方 法

本研究では，樹木が存在する画素と存在しない画素とで性質が大きく異なることを考慮し，評価方法を回帰タスクと分類タスクに分けて設計した．

まず，正解データにおいて樹冠高が正の値を持つ画素を「樹木が存在する画素」とみなし，これらに対しては樹冠高推定を回帰問題として評価する．回帰性能の指標として，平均絶対誤差（MAE），平均二乗誤差（MSE），二乗平均平方根誤差（RMSE）を用いた．一方で，正解データにおいて樹冠高が 0 である画素を「樹木が存在しない画素」とみなし，全画素を対象として樹木の有無を判定する 2 値分類問題として評価を行った．テストデータにおける樹冠高データセットの画素分布は， 3.2 節の表 1 に示す通り，樹木が存在する画素が 21,156,985 個，樹木が存在しない画素が 44,379,015 個である．正解データの樹冠高は整数型として与えられる一方，モデルの出力は小数点型であるため，分類評価において閾値を用いた処理を適用した．具体的には，モデルの出力値が閾値以上の場合を「樹木あり」，閾値未満の場合を「樹木なし」と判定した．本研究では，予測値が 0.5 未満の場合に当該画素を「樹木なし」と判定する閾値を設定した．2 値分類の性能評価には，混同行列に基づく正解率，再現率，適合率，F 値を用いた．本研究では，再現率は樹木が存在する画素をどれだけ検出できたかを示す．適合率は樹木が存在すると予測した画素のうち，実際に樹木が存在していた割合を示す．この評価方法により，樹木が存在する領域における

表 2: 樹木存在画素における各手法の回帰性能比較

	MAE	MSE	RMSE
U-Net (3ch)	5.25	49.05	7.00
Bottleneck Concat	4.48	36.50	6.04
Cross-Attn (Bottleneck)	4.42	36.71	6.06
Cross-Attn (ALL Layer)	**3.99**	**29.74**	**5.45**

表 3: 各手法の樹木有無の分類性能比較

	正解率	再現率	適合率	F 値
U-Net (3ch)	0.868	0.701	**0.864**	0.774
Bottleneck Concat	**0.876**	0.763	0.839	0.799
Cross-Attn (Bottleneck)	0.873	0.828	0.778	**0.808**
Cross-Attn (ALL Layer)	0.780	**0.956**	0.600	0.737

樹冠高の推定性能と，樹木の有無の分類性能を評価できる．

5.3 実験結果

本節では，4 節で示した 4 つの手法を，5.2 節で示した 2 通りの方法で評価する．

5.3.1 樹木存在画素における樹冠高推定性能

樹木存在画素における各手法の樹冠高推定性能の比較結果を表 2 に示す．表 2 に着目すると，いずれの評価指標においても Cross-Attn（ALL Layer）が最も高い性能を示していることが確認できる．特に，ベースラインである U-Net（3ch）との比較では大幅な誤差の低減が確認された．これは，エンコーダからボトルネック，デコーダに至るまで各層に Cross-Attention を導入することで，各解像度において地理情報と画像特徴の関係を段階的に学習できたためであると考えられる．また，Bottleneck Concat および Cross-Attn（Bottleneck）も U-Net（3ch）と比較して一定の性能向上を示している．

5.3.2 樹木有無判定における分類性能評価

各手法の樹木有無の分類性能の比較結果を表 3 に示す．

表 3 に着目すると，各手法の分類性能には明確な傾向の違いが確認できる．まず，正解率に着目すると，Bottleneck Concat が 0.876 と最も高く，樹木有無の分類性能に優れていることが分かる．また，U-Net（3ch）と Cross-Attn（Bottleneck）も 0.87 前後の正解率を示しており，大きな差は見られない．しかしながら，5.2 節に示す通り，本研究で扱った CHM データセットは，樹木が存在しない画素が全体の約 7 割を占めており，クラス分布に偏りが存在する．このような不均衡データにおいては，負例を正しく分類するだけでも高い正解率が得られるため，正解率のみではモデルの実質的な分類性能を適切に評価できない可能性がある．そこで，F 値に着目すると，Cross-Attn（Bottleneck）が 0.808 と最も高い値を示しており，検出性能と分類安定性のバランスに優れていることが分かる．次いで，Bottleneck Concat も 0.799 と高い値を示しており，ボトルネック層において画像特徴と地理情報を統合する手法が樹木有無の分類に有効であることが確認された．一方で，Cross-Attn（ALL Layer）は再現率が 0.956 と極めて高いが，適合率が 0.600 と低く，樹木が存在しない画素を誤って樹木が存在すると判定する傾向が強いことを示している．この結果として，正解率および F 値は他手法と比較して低下している．

5.4 考　　察

本節では，1 節で述べた 3 つのリサーチクエスチョン（RQ）に対する考察を行う．

5.4.1 RQ1 衛星画像からの樹冠高の推定性能

RQ1 は，衛星画像から樹冠高をどの程度推定できるかというものであった．表 2 より，衛星画像のみを入力とする U-Net（3ch）において，MAE=5.25, MSE=49.05，RMSE=7.00 となっている．これは，平均して 5,25m の誤差で樹冠高を推定していることを示している．

5.4.2 RQ2 地理情報の有効性

RQ2 は，地理情報を考慮することで樹冠高の推定性能は向上するかというものであった．表 2 より，すべての地理情報統合手法において U-Net（3ch）との比較で推定性能の向上が見られる．この結果から，地理情報を用いることで衛星画像のみでは捉えにくい環境条件を考慮でき，樹冠高の推定性能の向上に寄与することが示された．したがって，地理情報は，樹冠高の推定に有効であると結論付けられる．

5.4.3 RQ3 地理情報の統合手法による推定性能の違い

RQ3 は，樹冠高の推定において，地理情報を効果的に考慮するにはどのような方法が最適かというものであった．本研究では，Bottleneck Concat，Cross-Attn（Bottleneck），Cross-Attn（ALL Layer）の 3 手法を比較した．

まず，回帰性能に着目すると，Cross-Attn（ALL Layer）が最も高い性能を示している．これは，複数の異なる解像度に対して地理情報を段階的に統合する構成が，樹冠高の推定に有効であることを示している．続いて，分類性能に着目すると，Cross-Attn（Bottleneck）および Bottleneck Concat が比較的高い F 値を示しており，検出性能と分類安定性のバランスに優れていることが確認された．これらの結果から，地理情報を全層に導入する手法は樹冠高を推定するタスクには優れているが，樹木の存在有無の分類では誤検出が増加する傾向があり，分類安定性の観点では課題が残ることが分かった．

これは，Cross-Attention を全層に適用した場合は，各解像度において地理情報が反映され，地理条件を詳細に学習できる．しかしながら，局所的な画像特徴が過度に補正されるため，樹木の非存在領域においても樹木が存在すると誤認識しやすくなるためであると考えられる．その結果，樹冠高の推定性能は向上するものの，樹木有無の誤検出の増加により分類性能が低下したと考えられる．一方で，ボトルネック層のみに地理情報を統合する場合は，抽象的な特徴に対してのみ地理情報を付与するため，画像本来の局所的構造や境界情報が保持されやすい．このため，樹木の有無の判別においては安定した判断が可能となり，分類性能の向上につながったと考えられる．

6 まとめと今後の課題

本研究では，衛星画像を入力として，各画素に対応する樹冠

高を推定する手法を提案した．さらに，地理情報を統合する手法として 3 種類の手法を提案し，推定性能の比較を行った。実験の結果，地理情報を考慮することで樹冠高の推定性能が向上することを確認した．特に，Cross-Attention を全層に導入する手法が樹冠高の推定において最も高い性能を示した．

今後の課題として，対象地域を拡張し，未知地域における樹冠高と樹木有無を高精度で推定する手法の提案に着手する．

謝　　辞

本研究は，JSPS 科研費 JP25K03229，JP25K03228，JP24K03228 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Esmaeel Adrah, Wan Shafrina Wan Mohd Jaafar, Hamdan Omar, Shaurya Bajaj, Rodrigo Vieira Leite, Siti Munirah Mazlan, Carlos Alberto Silva, Maggie Chel Gee Ooi, Mohd Nizam Mohd Said, Khairul Nizam Abdul Maulud, Afonso Cardil, and Mikey Mohan. Analyzing Canopy Height Patterns and Environmental Landscape Drivers in Tropical Forests Using NASA's GEDI Spaceborne LiDAR. *Remote Sensing*, Vol. 14, No. 13, pp. 1–21, Article 3172, 2022.

[2] Aitor Ameztegui, Marcos Rodrigues, Pere Joan Gelabert, Bernat Lavaquiol, and Lluis Coll. Maximum Height of Mountain Forests Abruptly Decreases above an Elevation Breakpoint. *GIScience & Remote Sensing*, Vol. 58, No. 3, pp. 442–454, 2021.

[3] Anıl Can Birdal, Uğur Avdan, and Tarık Türk. Estimating Tree Heights with Images from an Unmanned Aerial Vehicle. *Geomatics, Natural Hazards and Risk*, Vol. 8, No. 2, pp. 1144–1156, 2017.

[4] Alexey Dosovitskiy, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, and Houlsby Neil. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, pp. 1–21, 2021.

[5] Vincent Dumoulin and Francesco Visin. A Guide to Convolution Arithmetic for Deep Learning. *arXiv preprint arXiv:1603.07285*, 2016.

[6] Geoffrey A. Fricker, Nicholas W. Synes, Josep M. Serra-Diaz, Malcolm P. North, Frank W. Davis, and Janet Franklin. More than Climate? Predictors of Tree Canopy Height Vary with Scale in Complex Terrain, Sierra Nevada, CA (USA). *Forest Ecology and Management*, Vol. 434, pp. 142–153, 2019.

[7] Pere J. Gelabert, Marcos Rodrigues, Lluís Coll, Cristina Vega-García, and Aitor Améztegui. Maximum Tree Height in European Mountains Decreases above a Climate-Related Elevation Threshold. *Communications Earth & Environment*, Vol. 5, No. 1, pp. 1–9, Article 84, 2024.

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626*, 2022.

[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Dawei Ma, Qi Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pp. 3146–3154, 2017.

[10] Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. Cross-Attention Head Position Patterns can Align with Human Visual Concepts in Text-to-Image Generative Models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, 2024.

[11] Ehsan Rahimi, Pinliang Dong, and Chuleui Jung. Global Variations in the Relationship between Tree Canopy Height and Land Surface Temperature. *Tropical Ecology*, Vol. 66, pp. 496–506, 2025.

[12] Robin Rombach, Blattmann Andreas, Lorenz Dominik, Esser Patrick, and Ommer Björn. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the 40th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pp. 10684–10695, 2022.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, pp. 234–241, 2015.

[14] Glenn Slade, Karen Anderson, Hugh A. Graham, and Andrew M Cunliffe. Repeated Drone Photogrammetry Surveys Demonstrate that Reconstructed Canopy Heights are Sensitive to Wind Speed but Relatively Insensitive to Illumination Conditions. *International Journal of Remote Sensing*, Vol. 46, No. 1, pp. 24–41, 2025.

[15] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pp. 5644–5659, 2023.

[16] Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V. Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, Theo Moutakanni, Piotr Bojanowski, Tracy Johns, Brian White, Tobias Tiecke, and Camille Couprie. Very High Resolution Canopy Height Maps from RGB Imagery using Self-Supervised Vision Transformer and Convolutional Decoder Trained on Aerial LiDAR. *Remote Sensing of Environment*, Vol. 300, pp. 1–37 Article 113888, 2024.

[17] Giuseppina Vacca and Enrica Vecchi. UAV Photogrammetric Surveys for Tree Height Estimation. *Drones*, Vol. 8, No. 3, pp. 1–14 Article 106, 2024.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 5998–6008, 2017.

[19] Fabien H. Wagner, Sophia Roberts, Alison L. Ritz, Griffin Carter, Ricardo Dalagnol, Samuel Favrichon, Mayumi CM Hirye, Martin Brandt, Philippe Ciais, and Sassan Saatchi. Sub-Meter Tree Height Mapping of California using Aerial Images and LiDAR-Informed U-Net Model. *Remote Sensing of Environment*, Vol. 305, pp. 1–13, Article 114099, 2024.

[20] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segmentation using Diffusion Models. In *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision (ICCV 2023)*, pp. 1206–1217, 2023.