DeepL Proに登録すると、より大きなサイズの文書ファイルを翻訳できます。
詳しくは、www.DeepL.com/pro をご覧ください。

arXiv:submit/5488358 [cs.CV] 2024年3月23

正確なガイダンスなしに学習する：低解像度の過去ラベルから大規模高解像度土地被覆マップを更新する

Zhuohong Li[1*] , Wei He[1*] , Jiepan Li[1] , Fangxiao Lu[1] , Hongyan Zhang[1,2†]
[1]武漢大学 [2]中国地球科学大学
{ashelee, weihe1990, jiepanli, fangxiaolu}@whu.edu.cn, zhanghongyan@cug.edu.cn

要旨

大規模な高解像度（HR）土地被覆マッピングは、地球表面を調査し、人類が直面する多くの課題を解決するために不可欠な作業である。しかし、複雑な地形、様々な地形、そして広範囲に渡る正確な学習ラベルの不足により、この作業は依然として容易ではない。本論文では、低解像度（LR）の過去の土地被覆データに容易にアクセスでき、大規模なHR土地被覆マッピングを導く効率的な弱教師付きフレームワーク（Paraformer）を提案する。具体的には、既存の土地被覆マッピングアプローチは、局所的な地表の詳細を保存するCNNの優位性を再証明しているが、様々な地形における大域的なモデリングがまだ不十分であるという問題を抱えている。そのため、我々はParaformerにおいて、ダウンサンプリングのないCNNブランチとTransformerブランチからなる並列CNN-Transformer特徴抽出器を設計し、局所的な文脈情報と大域的な文脈情報を共同で捕捉する。さらに、学習データの空間的不一致に直面し、擬似ラベル支援学習（PLAT）モジュールを採用し、HR画像の弱い教師付きセマンティックセグメンテーションのためのLRラベルを合理的に洗練する。2つの大規模データセットを用いた実験により、ParaformerがLR履歴ラベルからHR土地被覆マップを自動的にアップデイトする上で、他の最新手法よりも優れていることが実証された。

1. はじめに

土地被覆マッピングは、リモートセンシング画像の各画素に「耕作地」や「建物」などの土地被覆クラスを与えるセマンティック・セグメンテーション・タスクである[14]。土地被覆データは、自然や人間の活動によって頻繁に景観が変化するため、継続的に更新される必要がある[37]。センサーや衛星が発達するにつれて、大規模な高解像度（HR）リモートセンシング画像（1m/ピクセル以下）が容易に得られるようになった[1]。迅速で大規模なHR土地被覆マップ

*同等の貢献を示す。†共著者。コードとデータはhttps://github.com/LiZhuoHong/ Paraformerで公開されている。

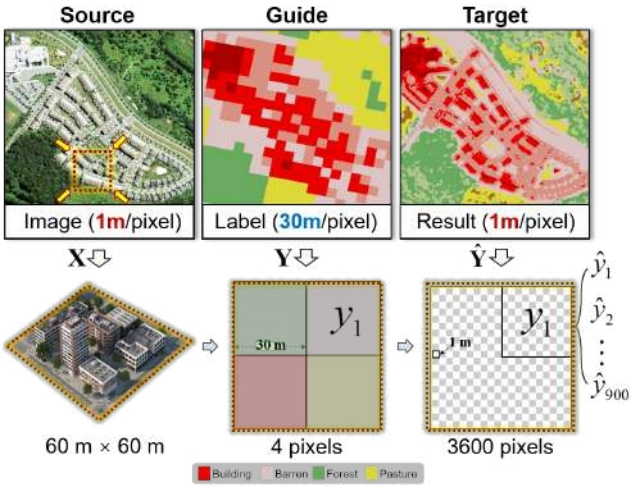

図1.HRリモートセンシング画像(**Source**)とLR履歴ラベル(**Guide**)を使用してHR土地被覆結果(**Target**)を生成する 際の解像度の不一致の問題。

最新のHR土地被覆データは、地表面を正確に描写できるため[21, 27, 55]、下流のアプリケーションを容易にするために、この更新はさらに重要である。しかし、HR画像に映し出される複雑な地表の詳細や、広域にわたる様々な地形は、大規模なHR土地被覆マップを定期的に更新する上で、依然として課題となっている[28]。

HR 画像の土地被覆マッピングのための先進的な手法は、長年にわたり畳み込みニューラルネットワー ク（CNN）によって支配されてきた。CNNベースの手法は、HR画 像 の セマンティック・セグメンテーションのための局所的な詳細を細かく捉えることができるが、畳み込み演算の本質的な局所性により、より広い領域にわたる様々な地形への実装にはまだ限界がある[2]。近年、Transformerは、セマンティックセグメンテーション[5, 18, 34]や、地球観測の大規模な応用において大きな成功を収めている。[11, 41, 48].CNNは、グローバルなコンテクストをモデル化するためにマルチヘッド自己注 意メカニズムを採用するが、低レベルの特徴が不足しているため、ローカルな詳細の表 現に苦労している[10, 48]。

さらに、CNNまたはTransformer構造を持つ現在の手法は、一般に、完全教師あり 戦略を採用することで、十分な元行動訓練ラベルに依存している。

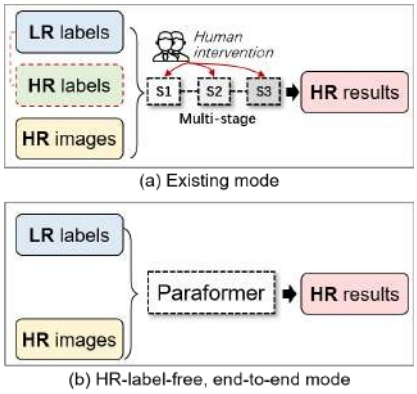

図2.LRラベルを用いた大規模なHR土地被覆マッピングの2つのモード。(a) 既存のモードは、部分的なHRラベルに答えるか、人間が介在するエンド・ツー・エンドの学習を必要とする。(b)**Paraformerは、HR**ラベルを使わず、エンド・ツー・エンドの訓練が可能なモードの形成を目指している。

[20, 32, 39].しかし、大規模な地理的領域に対して正確なHR土地被覆ラベルを作成することは、非常に時間と手間がかかる[6, 37]。

幸いなことに、過去数十年の間に、広い範囲をカバーする多くの低解像度（LR）土地被覆データが既に出現している[9, 22, 44, 56]。これらのLR土地被覆履歴データを代替ガイダンスとして利用することは、HRラベルの不足を緩和する方法である[29]。とはいえ、HR画像とLRラベルの学習ペアが一致していないため、完全教師あり手法には課題があった。さらに、適用されるシナリオが異なるため、自然シーンに対する既存の弱い教師ありセマンティックセグメンテーション手法（例えば、バウンディングボックスや画像レベルのラベルからの学習）は、同様にこの課題を処理するのに適用できない[15, 23, 24, 57]。特徴的なのは、LR土地被覆の不正確なサンプルは、地球観測中に衛星から異なる空間解像度でもたらされることである。図1に示すように、HR（1m/pixel）画像**Xからは**、60m×*60*mの領域にある物体が明瞭に観察できるが、LR（30m/pixel）ラベル**Yでは**、その領域は4画素でしかラベル付けされていない。そのため1-m土地被覆の結果$\mathbf{Y}$は、ラベル付けされたピクセルy_1。900 個の対象画素$\{\hat{y}_1, \hat{y}_2 - - \hat{y}_{900}\}$のガイド情報は、地理空間的なミスマッチを引き起こす。大規模なHR衛星画像のセマンティックセグメンテーションのための唯一のガイダンスとしてLRラベルをどのように再利用するかは、地球観測とコンピュータビジョンの分野で共有されている特別な問題である[28, 31, 37]。大規模なHR土地被覆マッピングのためにLRラベルを利用する最新の方法を要約すると、まだ2つの主要な問題がある：

1. *広範な応用分野に対して、既存の特徴抽出トラクターは、HR画像から局所的な詳細を捉え、様々な土地形態におけるグローバルなコンテキストを一度にモデル化することは困難である[29, 54]。*
2. *訓練ペアのミスマッチに対して、図2 (a)に示すような既存のパイプラインは、依然として部分的なHRラベルに依存しているか、人間が介入して非エンド・ツー・エンドの最適化を必要とする[12, 27]。*

これらの問題を解決するために、図2(b)に示すように、我々はLR土地被覆ラベルを用いた大規模なHR土地被覆マッピングを導く、HRラベルフリーのエンドツーエンドフレームワークとしてParaformerを提案する。具体的には、Paraformerはダウンサンプリングフリーの CNNブランチとTransformerブランチを並列にハイブリッドし、大規模なHR画像からローカルとグローバルのコンテクストを共同でキャプチャし、擬似ラベル支援学習（PLAT）モジュールを採用して、フレームワーク学習のためにLRラベルから信頼性の高いインフォーメーションを掘り起こす。

本研究の主な貢献は以下の通りである：(**b)**ダウンサンプリングフリーのCNNブランチをTransformerブランチと並列にハイブリッド化し、高い空間分解能と深いレベルの表現の両方を持つ特徴を捉える。**(c)**PLATモジュールは、フレームワーク学習を導くために、ラベル付けされたサンプルを常に改良するために、原始予測とLRラ ベルを反復的に交差させる。これは、LR履歴データから大規模なHR土地被覆マップを更新する簡潔な方法を提供する。

2. 関連作品

土地被覆マッピングのアプローチ初期の段階では、決定木[19]、ランダムフォレスト[7]、サポートベクトルマシン[40]などのピクセル間分類法が、マルチスペクトルLR画像の土地被覆マッピングによく用いられた。しかし、これらの手法は、一般的にコンテ クス情報を無視し、光学的なHR画像には空間的な詳細情報は多いが、 スペクトル的な特徴は限られているため、HRの場合、断片的な結果しか得 られない[29]。データ駆動型セマンティック・セグメンテーションの発展 に伴い、多くのCNNベースの手法が、HR画像の土地被覆マッ ピングに広く使われるようになった[37, 52, 53]。さらに、別のアーキテクチャとして、Transformer は、シーケンス間モデリング[3, 10, 30]でグローバルなコン テキストをキャプチャする際に大きな力を発揮し、建物抽出[25, 41]、道路検出[11]、土地オブジェクト分類[47]などの地球観測の多くの大規模なアプリケーションで卓越した性能を示す。さらに、セグメント何でもモデル（SAM）[35, 50]を用いて、より細かいラベルを生成するための省力化によって新しい方法を開発した研究も多い。しかし、CNNベースとTransformerベースの両方の手法の大規模な応用の基礎となるのは、十分な正確な学習ラベルである。HRラベルの乏しさは、これらの完全教師ありアプローチによる大規模なHR土地被覆マッピングの妨げとなっている。

土地被覆ラベル付きデータ： 手作業や半手動のアノテーションによって大規模なHRラベルを作成するのは、非常に時間とコストがかかる[17, 36]。そのため、既存のHR土地被覆データは一般に小規模なものに限られている。例えば、LoveDAデータセットには、0.3mの土地被覆データが含まれており、次のような範囲をカバーしている。

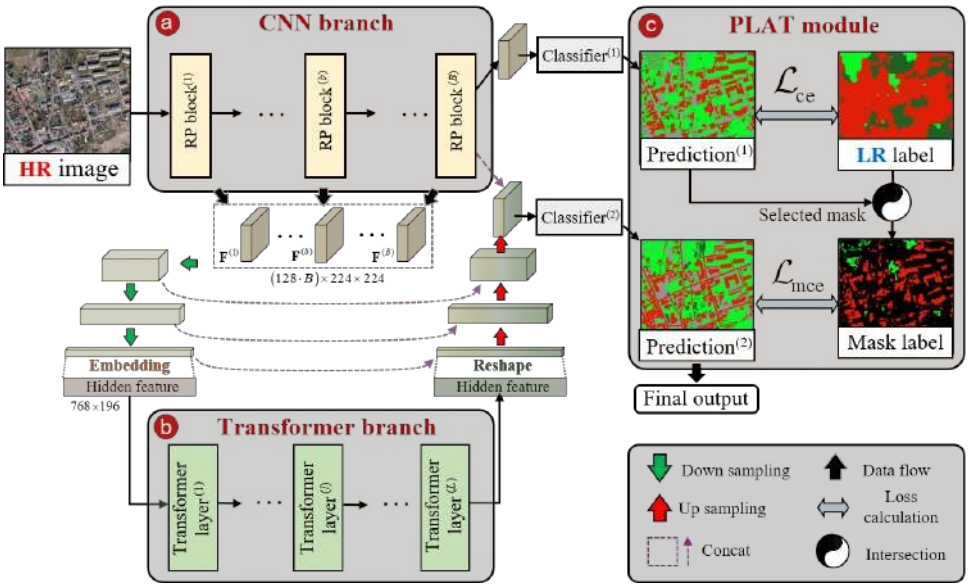

図3.Paraformerの全体的なワークフロー。このフレームワークはHR画像とLRラベルのみを訓練入力とし、3つのコンポーネントを含む：(a)CNNベースの解像度保存ブランチ、(b)Transformerベースのグローバルモデリングブランチ、(c)擬似ラベル支援学習(PLAT)モジュールである。

中国の536.15 km^2 [46]。Agri-visionデータセットは、0.1mのラベル付きデータを含み、米国の560 km^2 をカバーしている[13]。契約では、一般にLR土地被覆データはより広い範囲をカバーしている。例えば、米国地質調査所（United States Geological Survey）は、米国全土をカバーする30mの土地被覆データを周期的に更新している[49]。欧州宇宙機関(ESA)は、2020年以降、世界の10m地表面データを毎年更新している[44]。これらのLRデータは、大規模なHR土地被覆マッピングを導くための代替ラベルソースとみなすことができる。しかし、膨大で正確でない標識サンプルは、まだ実用化の妨げとなっている。

LR履歴ラベルマイニングの戦略大規模なHR土地被覆マッピングにおける正確なラベルの不足を緩和するために、多くの研究がLRラベルから信頼できる情報をマイニングする努力を行ってきた。例えば、ラベル超解像ネットワークは、HRラベルから推測される統計的分布を利用して、LRラベルの不正確な部分を制約するために設計された[31, 37]。WESUPと名付けられた多段階フレームワークは、30mラベルを用いた10m土地被覆マッピングのために構築された[12]。WESUPでは、LRラベルからきれいなサンプルを絞り込むために、マルチモデルが学習された。同様に、2021年IEEE GRSSデータ融合コンテスト（DFC）の優勝者のアプローチは、30-mラベルを洗練するために浅いCNNを採用し、その後、米国メリーランド州の1-m土地被覆マップを作成するために擬似ラベルでマルチモデルを学習した[27]。さらに、弱教師付き損失関数によって、LRラベルの確実な部分を選択する低-高ネットワーク（L2HNet）が提案された[28]。利用可能な10mのラベルで中国全土の1mの土地被覆マップを作成するために、7つのL2HNetが広い地理的領域に適応するように選択的に訓練された[29]。

パファロマーは、部分的なHRラベルに頼るか、人間の介入を必要とするこれらのアプローチとは異なる。

は、大規模なHR土地被覆マッピングを容易にする、HRラベルフリーのエンドツーエンドフレームワークとして設計されている。

3. 方法論

Paraformerは、大規模なHR土地被覆マッピングのために、ローカルとグローバルのコンテクストを共同で捉え、LRラベルを合理的に利用するために、並列CNNとTransformerのブランチとPLATモジュールを組み合わせている。本節では3つの構成要素を順次紹介する。

3.1. CNNベースの解像度保存分岐

Paraformerの基本的な特徴抽出器として、また以前のL2HNet V1[28]の主要な構造として、CNNブランチはHR画像から局所的なコンテキストを捉え、特徴のダウンサンプリングを防ぐことで空間的な詳細を保存するように設計されている。図3(a)に示すように、CNNブランチは5つの直列接続された解像度保存(RP)ブロックによって構成される。各RPブロックは1×1、3×3、5×5のサイズの並列畳み込み層を含み、そのステップは特徴サイズ維持のために1に設定される。インセプションモジュール[42]と部分的に類似しているが、各ブロック内の異なるスケールの層のチャンネル数は、それらのカーネルサイズに反比例しており、128、64、および5×5に設定されている。

32.この設定に基づき、RPブロックは、深いエンコーダー・デコーダーパターンで特徴マップをダウンサンプリングする代わりに、適切な受容野で特徴を捉えることができる。直列ブロックは、1×1カーネルを多用することで、特徴量の空間的な変化を十分に保持することを目的としている。3×3と5×5のカーネルは必要な周辺情報を捕捉する。さらに、マルチスケール特徴マップは連結され、128チャンネルに縮小される。

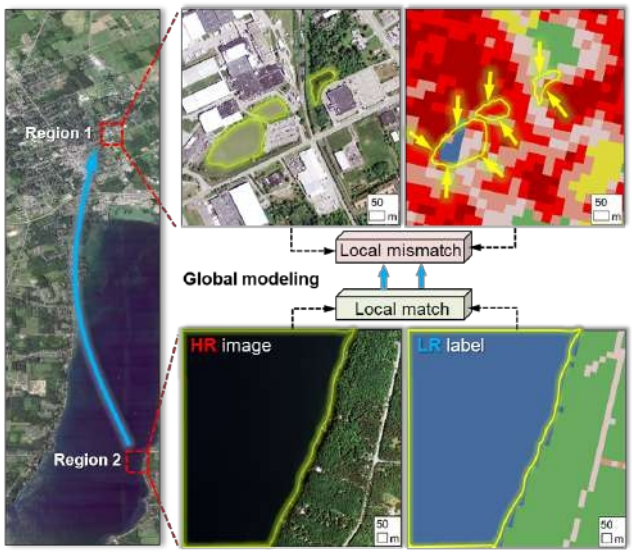

図4.2つの領域における局所的なミスマッチ／マッチの例。水辺は黄色の境界線で示されている。リージョン1は都市部周辺に散在する湖で、アノテーションは一致しない。領域2は一致したアノテーションを持つ大規模な河川。

エニング。また、残差学習と細部保存のために、ブロック間のショートカット接続が採用されている。

3.2. トランスフォーマーベースのグローバル・モデリング・ブランチ

同じ土地被覆クラスであっても、HR画像では特徴的な属性を持ち、LRラベルでは異なる注釈が付されている場合がある。図4は
さまざまな地域にある湖や川を考慮することで
本質的な局所性を持つCNNブランチが、CNNブランチを阻害している。

CNNブランチをTransformerブランチとさらにハイブリッド化することで、大規模エリアにおける様々な地形への適応を可能にする。

Transformer」ブランチは、グローバルな文脈をとらえ、分散した地域間で長期的なサポートを構築することを目的としている。図3（b）に示すように、トランスフォーマー・ブランチには12個のトランスフォーマーが含まれている。

以前のレイヤー。 各レイヤーはレイヤーの正規化を含む、
多頭自己注意、多層知覚。そして

作業トレーニングPLATモジュールは、特定できないサンプルを選別し、LRラベルから信頼できる情報を掘り起こすことを目的としている。具体的には、PLATモジュールの2つの部分を以下に説明する。CNNブランチでは、3×3畳み込みレイ ヤーで構成される分類器(1) を用いて、HR特徴マップに基づく原始予測(1) を生成する。そして、$\hat{\mathbf{Y}}$ として表現される予測値(1) と、**Yとして**表現されるLRラベルとの間のクロスエントロピー（CE）損失を計算する。形式的には、H、W、Lをパッチの高さ、重さ、土地被覆クラスとみなすことで、CNNブランチのCE損失は次のように書かれる：

$$\text{ルース}(\mathbf{Y}, \tilde{\mathbf{Y}}) = \frac{\sum_{i=0}^{W}\sum_{j=0}^{H}\sum_{l=1}^{L} y_{ij}^{(l)} \log(\hat{y}_{ij}^{(l)})}{H \times W}. \quad (1)$$

フレームワークの最終出力として、予測(2) は、CNNとTransformerブランチの連結された特徴マップから分類され、$\hat{\mathbf{Y}}'$ として表される。各訓練反復の間、マスク・ラベルを生成するために、予測(1) と LRラベルの単純だが効果的な**交差を**取る。具体的には、マスクラベル中の矛盾したサンプルは、損失計算から除外するために空値として設定される。さらに、CNNブランチの予測は、イメージ年齢と整合性の高いHRテキスト情報を含むため、マスクラベルは細かいエッジの輪郭を描き、ラベル付けされたサンプルを保持する。最後に、提案されたマスク・クロス・エントロピー（MCE）損失が予測(2) とマスク・ラベルの間で計算される。正式には、MCE損失は次のように書かれる：

$$\text{Lmce}(\mathbf{M} \cdot \mathbf{Y}, \tilde{\mathbf{Y}}) = \frac{\sum_{i=0}^{W}\sum_{j=0}^{H}\sum_{l=1}^{L} y_{ij}^{(l)} m_{ij} \log(\hat{y}_{ij}^{(l)})}{\text{Sum}(\mathbf{M}(i,j)=1)}. \quad (2)$$

式2において、MはH×Hの大きさの**交差**マスクである。

$W \cdot m_{ij}$, i∈[0, H] , j∈[0, W]は$\mathbf{M}(i,j)$の要素である。

と簡単に表すことができる：

$$m_{ij} = 1 / y_{ij} = Y_{ij}' \quad (3)$$

各RPブロックで抽出された特徴マップは連結され、Transformer分岐に入力される。具体的には、CNNブランチから抽出された特徴量はダウンサンプリングされ、隠れ特徴層に埋め込まれる。そして、Transformer分岐は、グローバルコンテキストをキャプチャするために、高密度の特徴パッチをエンコードする。その後、符号化された特徴量は常にHR画像のサイズにアップサンプリングされ、最終結果に分類される。アップサンプリングプロセスの間、各ステージで出力された特徴量はエンコード前の特徴量と連結され、最終的な特徴マップに膨大な局所コンテキスト情報をもたらします。

3.3. 疑似ラベル支援トレーニング・モジュール

図3(c)に示すように、弱いLRラベルで大規模なHR土地被覆マッピングを合理的にガイドするために、弱教師付きPLATモジュールがフレームを最適化するために採用されている。

$$0 | \, {}_{Yij} \models \acute{Y} \, ij.$$

パラフォーマーの全損失は、2つのブランチの損失の組み合わせであり、次のように書かれる：

$$\text{Ltotal} = \text{Lce} + \text{Lmce}. \tag{4}$$

4. 実験

4.1. 調査地域とデータの利用

様々な土地形態と異なるLRラベルに対するParaformerを総合的に評価するため、2つの大規模データセットで実験を行った。

チェサピーク湾のデータセットは、米国最大の河口から採取され、732の重複しないタイルに編成されており、各タイルのサイズは6000×7500ピクセルである[37]。具体的なデータには以下が含まれる：

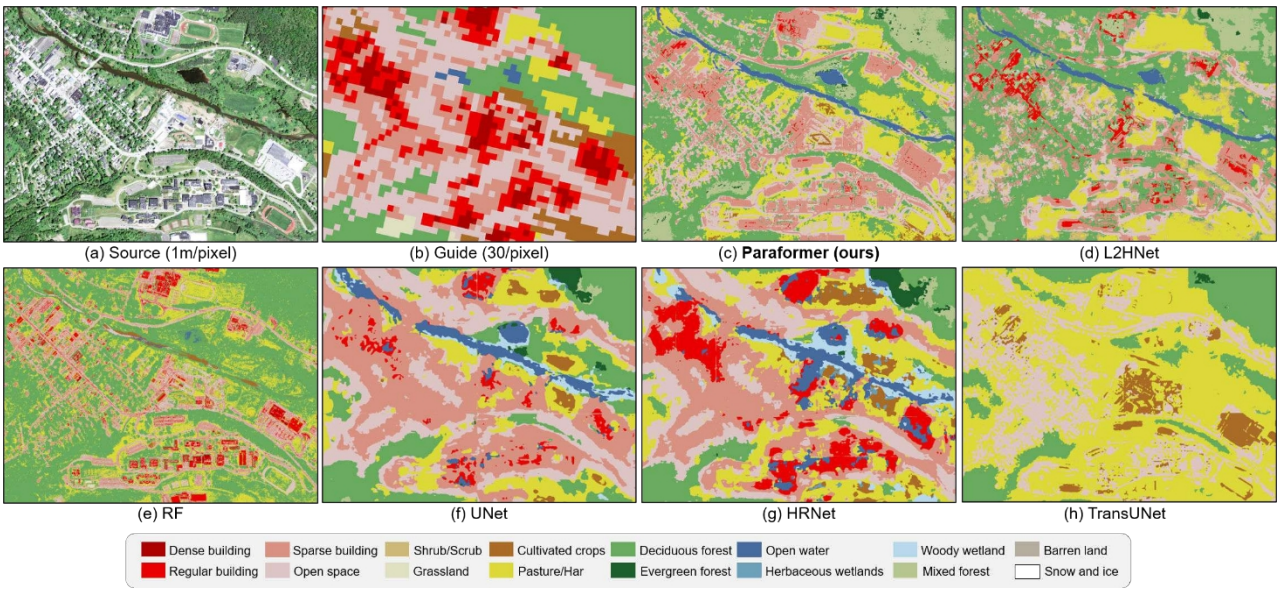

図5.16クラスからなるチェサピーク湾のデータセットにおける、トレーニングデータのデモと、Paraformerと他の代表的な手法の視覚的比較。(a) HR画像。(b) LRラベル。(c) Parafomerによる土地被覆マッピング結果。(d-h) 代表的な5つの手法による土地被覆マッピング結果。

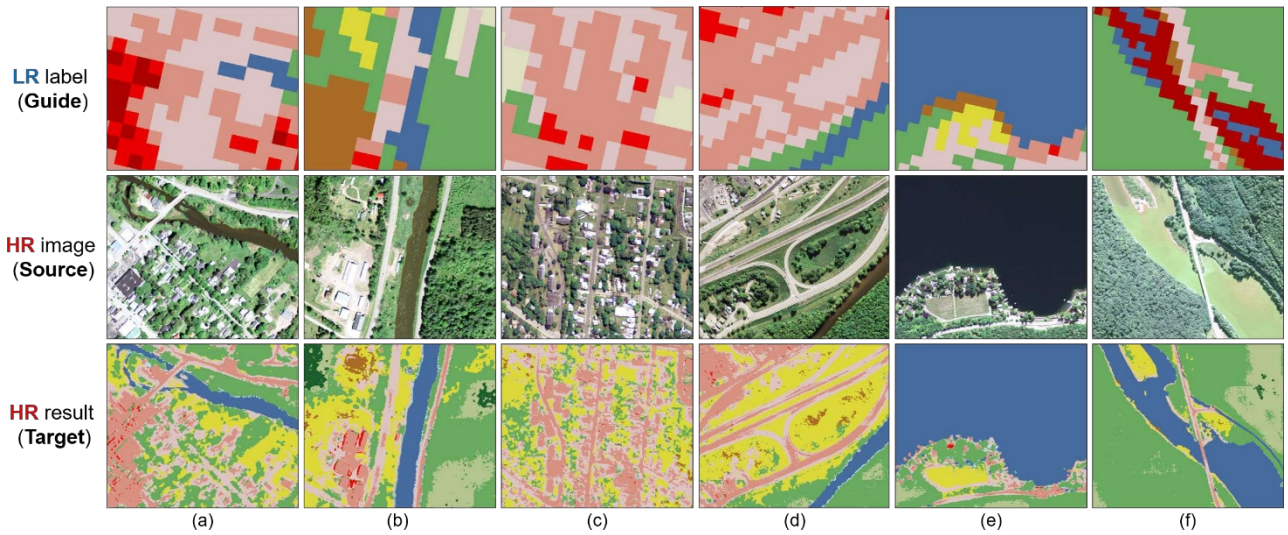

図6.チェサピーク湾のデータセットで観測スケールを細かくした6つの典型的なエリア。1行目はLRラベル**（ガイド）**。2行目はHR画像**（ソース）**。3行目は**Paraformerによって**生成されたHR結果**（Target）**。

1. *HR画像（1 m/ピクセル）は*、米国農務省の全米農業画像プログラム（NAIP）のものである。この画像には、赤、緑、青、近赤外 の4つのバンドが含まれている[33]。
2. *LR ヒストリカル ラベル（30m/ ピクセル）は*、USGSの国土被覆データベース（NLCD）[49]から取得したもので、16の土地被覆クラスを含む。
3. *地上絵（1m/ピクセル）は*CCLC（Chesapeake Bay Conservancy Land Cover）プロジェクトによるもの。

ポーランドのデータセットには、ポーランドの14の州が含まれ、各タイルのサイズは1024×1024ピクセルで、403の非重複タイルに編成されている。具体的なデータは以下の通り：

1. *HR画像(0.25m と0.5m/pixel)*は、以下のものである。

LandCover.ai [4]データセット。画像には赤、緑、青の3つのバンドが含まれている。

2. *LR履歴ラベルは*、3種類の10m土地被覆データと1種類の30mデータから収集され、それらはFROM GLC10 [9]、ESA GLC10 [44]、ESRI GLC10 [22]、GLC FCS30 [56]と名付けられた。
3. *HRグランドトゥルースは*、7つの土地被覆クラスを持つOpenEarthMap [51]データセットからのものである。

4.2. 実施の詳細と指標

実験では、全ての手法はLR土地被覆データのみをトレーニング・ラベルとして用いる。Paraformerはパッチサイズ224×224、バッチサイズ8のAdamWオプティマイザで学習される。

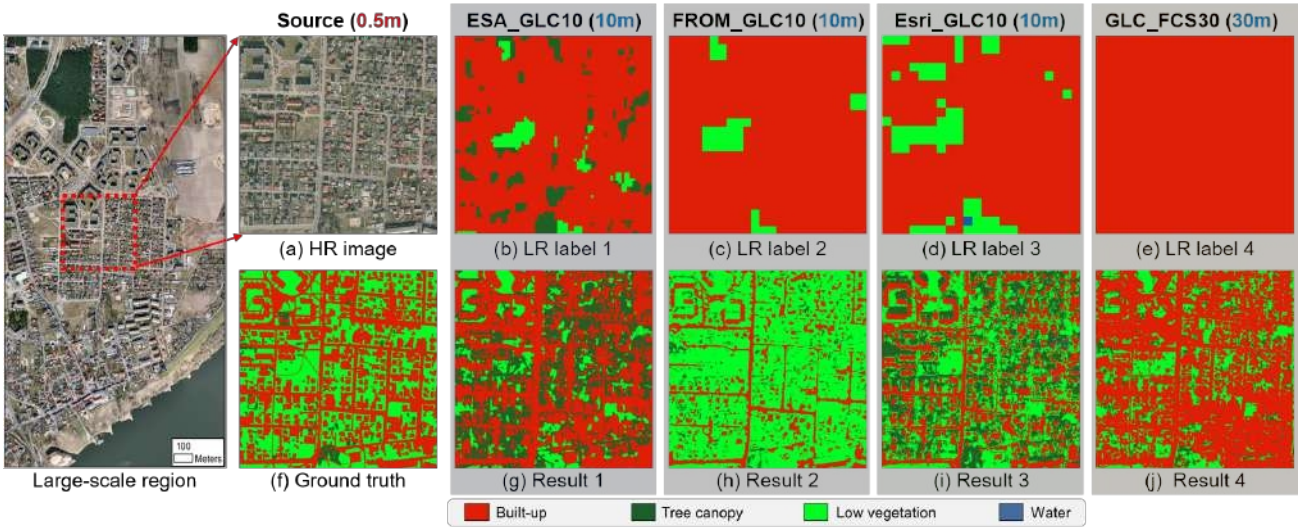

図7.ポーランドデータセットにおける**Paraformerの**視覚的結果。デモ領域は大規模訓練領域から抽出された訓練ピースの1つである。(a-e) HR画像(0.5 m/pixel)とESA GLC (10 m/pixel)、FROM GLC (10 m/pixel)、Esri GLC (10 m/pixel)、GLC FCS30 (30m/pixel)の4種類のLRラベルの学習ペア。(f-g)グランドトゥルース(0.5 m/pixel)と、異なるLRラベルを用いたParaformerのマッピング結果。

解像度 gap	方法	チェサピーク湾流域6州のmIoU(%)						
			デラウェア州	ニューヨーク州	メリーランド	バージニア	ウェストバージニア	ペンシルバニア
						平均		
30×	**パラフォーマー**	**65.57**	**71.43**	**70.20**	**60.04**	68.01	52.62	**64.65**
	L2HNet [28]	61.77	68.12	65.24	58.52	**69.39**	**55.43**	63.08
	TransUNet [10]	53.15	60.53	60.42	51.08	66.21	47.52	56.49
	ConViT [18]	55.26	60.71	61.58	53.94	59.80	49.11	56.73
	CoAtNet [16]	56.89	62.83	61.25	53.57	65.67	51.34	58.59
	モバイルヴィット[34]	58.03	61.32	61.84	55.53	57.04	48.64	57.07
	EfficientViT[5]	53.72	61.28	59.48	51.38	57.34	48.76	55.33
	ユネットフォーマー[48]	58.85	65.11	61.34	59.10	60.84	47.20	58.74
	DC-スウィン[47]	59.65	65.99	58.60	58.06	64.11	48.15	59.09
	ユネット[38]	54.16	58.79	56.42	53.21	57.34	46.11	54.34
	HRネット[45]	52.11	56.21	50.76	50.03	57.48	45.42	52.00
	リンクネット[8]	58.27	62.05	52.96	52.11	48.71	48.93	53.84
	スキップFCN [26]	60.97	64.83	59.44	55.37	64.72	54.66	60.00
	SSDA【43	57.91	61.54	54.85	51.71	57.71	47.15	55.15
	RF [7]	59.35	55.03	55.26	51.07	52.29	54.36	54.56

表1.チェサピーク湾流域の6つの州におけるParaformerと他の手法の定量的比較。すべての手法は1mの画像と30mのラベルで学習された。異なる手法のmIoU（%）は、その結果と1-m ground truthとの間で計算された。

最大ギャップ	LRラベル	各手法のmIoU(%)								
		パラフォーマー (私たちの)	L2Hネット [28]	トランズネット [10]	コンヴィット [18]	モバイルヴィット [34]	DC-スウィン [47]	HRネット [45]	スキップFCN [26]	RF [7]
40×	FROM_GLC10 [9]	**56.57**	50.15	38.44	39.36	41.03	43.56	43.66	27.14	21.48
	ESA GLC10 [44]	**55.19**	52.13	35.58	36.09	38.42	40.05	49.81	28.34	26.97
	Esri GLC10 [22]	**55.07**	50.78	37.79	38.78	38.50	39.91	46.65	28.18	19.36
120×	GLC FCS30 [56]	**49.39**	43.62	26.20	29.16	29.57	30.14	41.46	23.67	17.02

表2.ポーランドデータセットでの定量的比較。3種類の10mラベル（FROM GLC10、ESA GLC10、Esri GLC10）と1種類の30mラベル（GLC FCS30）で学習したParaformerと他の手法のmIoU（%）を示す。

8エポックにわたって損失が減少しなくなったとき。土地被覆クラスが4つの基本クラスに統一された後、その結果とHRグランドトゥルースとの間で、mean intersection over union (mIoU)の測定値が計算される。比較される手法は以下の通りである：ランダムフォレスト（RF）は、大規模な土地被覆マッピングで広く

使用されているピクセル間手法である[7]。TransUNet [10]、ConViT [18]、CoAtNet [16]、Mobile- ViT [34]、EfficientViT [5] は CNN-Transformer-Hy-Transformer である。

セマンティック・セグメンテーションのためのブリッジング手法。UNetformer[48]とDC-Swin[47]は、リモートセンシング画像専用のCNN変換メソッドである。UNet[38]、HRNet[45]、LinkNet[8]は典型的なCNNベースのセマンティック・セグメンテーション手法で、HRランドカバーマッピング[37, 52, 53]で広く採用されている。SkipFCN[26]とSSDA[43]は、30mラベルから1m土地被覆変化マップを更新するための浅いCNNベースの手法であり、1位と2位を獲得した。

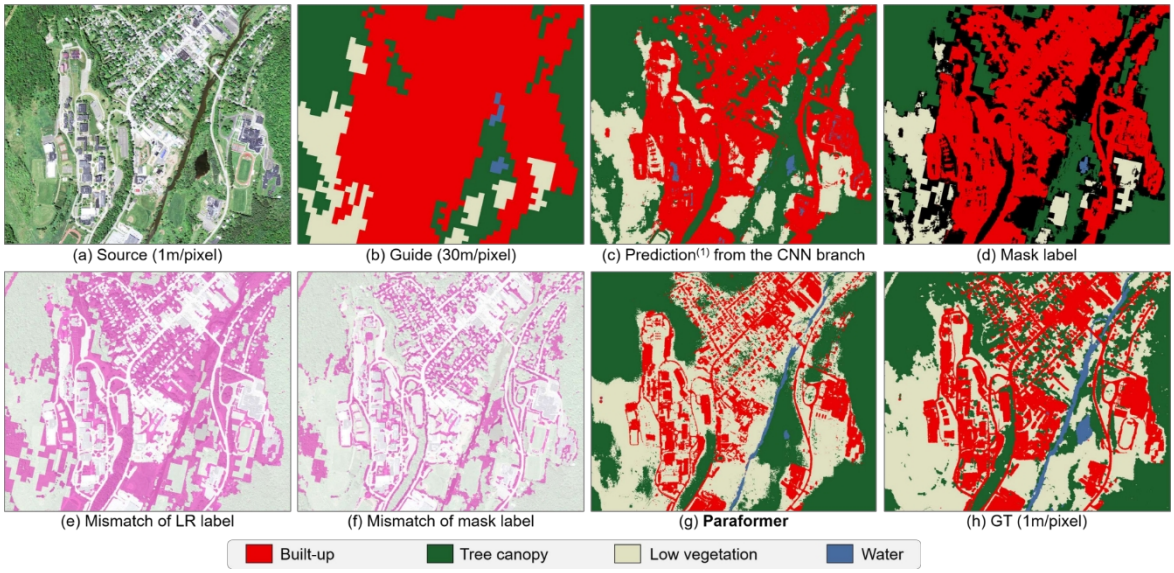

図8.チェサピーク湾のデータセットからサンプリングされた、4つの統一されたクラスを持つParaformerの学習データとさまざまな出力の例。(a) HR画像。
(b) LRラベル。(c)CNNブランチからの原始予測。(d) (b)と(c)の交点部分としてのマスク・ラベル。**黒い**部分は教師情報なしで無効とする。
(e-f)LRラベルとマスクラベルの不正サンプル（ピンク色）。(g) Paraformerの最終結果。(h) HR ground truth。

2021 IEEE GRSS DFC [27]で発表された。L2HNetは、弱教師付き土地被覆マッピングのために設計された最先端の手法である[28]。

4.3. 比較結果

チェサピーク湾データセットでの比較：表1と図5は、チェサピーク湾データセットでの比較を示している。定量的な結果から、Paraformerはデラウェア州、ニューヨーク州、メリーランド州、ペンシルベニア州で優れた結果を示した。L2HNetはヴァージニア州とウェストヴァージニア州で優れた結果を示した。平均すると、ParaformerのHR土地被覆マッ ピング結果は、エンタイヤ地域で最も正確で、mIoUは64.65%である。図5(c)に示すように、Paraformerの視覚的結果は、他の手法と比較してHR画像との整合性が高い。完全教師ありのセマンティックセグメンテーション課題とは異なり、一致しない訓練ペアはモデル訓練中に重大な見当違いを引き起こす可能性がある。例えば、図5(f)と(g)に示すような大まかな結果では、UNetとHRNetは特徴量を過剰にサンプリングし、HR画像と一致するのではなく、LRラベルに合うように結果を促している。さらに、定量的な結果から、UNet、LinkNet、HRNetの性能は不十分で、mIoUは54.34%、53.84%、52.00%であった。比較されたCNN-Transformer手法（例えばTransUNet）は、ローカルとグローバルのコンテクスト情報を組み合わせるが、その構造は特徴解像度の事前提供や地理空間的ミスマッチの処理に重点を置いていない。その結果、TransUNetは、図5 (h)に示すように、視覚的な結果では弱い性能を示し、mIoUは56.49%である。さらに、SkipFCN、SSDA、RFは、小さな受容野や画素間戦略を用いて、視覚的な不一致を解消する。

しかし、SkipFCN、SSDA、RFは、深いレベルの特徴表現と大域的な状況情報がないため、mIoUは59.99%、55.15%、54%となる。しかし、SkipFCNは59.99%、SSDAは55.15%、RFは54.56%である。図5(e)に示すように、RFは地表の詳細を精緻に 予測するが、河川、湖沼、牧草地を誤って分類する。異なる地形に対するParaformerの効果をさらに実証するために、6つの典型的な地域を図6に示す。視覚的な結果は、HR土地被覆マップの様々な地形間の複雑な地面の詳細が、LRの過去の土地被覆ラベルからうまく更新できることを示している。

ポーランドデータセットでの比較： ポーランドのデータセットを用いた実験では、4つのLRラベルを個別に利用し、ポーランドの14州の0.25/0.5mの土地被覆マップを作成するために、すべての手法が使用された。これらのLRラベルは、10-m FROM GLC10、ESA GLC10、Esri GLC10、および
30-m GLC FCS30。表2に示すように、Paraformerは、より極端な地理空間ミスマッチにおいて、代表的な8つの手法（すなわち、弱教師付き、CNN-Transformer、CNNベース、ピクセル間アプローチ）と比較される。最先端の手法と比較して、Paraformerは10mラベルを利用することで、mIoUが6.42%、3.06%、4.29%増加した。30-m ラベルを最大120× の分解能で解決すると、Paraformer の mIoU は 49.39%となり、L2HNet と比較して 5.77%増加する。典型的なCNNベースの手法の平均mIoUは10mケースで46.71%、30mケースで41.46%である。スキップFCNとRFは全ての手法の中で最も低いmIoUを示し、これは極端にマッチしない状況を扱うことの難しさを示している。さらに、4つのケースで示されたParaformerの定量的な結果から、提案されたframe

アブレーション法	チェサピーク湾流域の6州のmIoU(%)								
	デラウェア州	新着情報	メリーランド州		ペンシルバニア		バージニア州		ウェストバ
		ージニア	**平均**	パラムス	フロップ数				
パラフォーマー	**65.57**	**71.43**	**70.20**	**60.04**	**68.01**	**52.62**	**64.65**	109.4M	141.3G
単独CNN支店	59.57	67.87	64.30	53.86	65.26	50.01	60.15	4.5M	56.1G
単独変圧器分岐	53.15	60.53	60.42	51.08	66.22	47.52	56.49	96.9M	83.3G
PLATなしのハイブリッド	62.69	70.39	67.15	58.33	67.47	50.83	62.81	109.4M	141.3G

表3.チェサピーク湾流域の6つの州におけるParaformerのアブレーション結果。単独CNNブランチ、単独Transformerブランチ、PLATなしハイブリッドは、それぞれCNNブランチ、Transformerブランチ、PLATモジュールの寄与を調べることを目的としている。

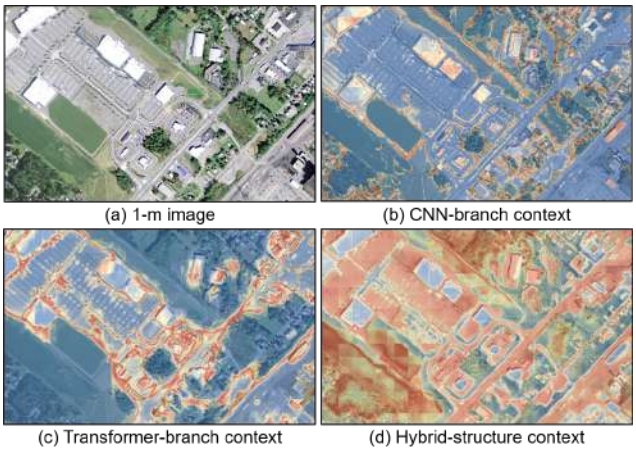

図9.アブレーションメソッドから抽出されたコンテクストのデモンストレーション。(a)元のHR画像。(b) 単独のCNN枝によって抽出されたコンテキスト。(d)CNN-変換器ハイブリッド・バックボーンによって抽出されたコンテキスト。

は、異なるLRラベルから安定した結果を得ている。図7は、4つのケースにおけるParaformerの視覚的な結果を示している。並列CNN-Transformer構造とPLATモジュールにより、Paraformerは、局所的に大まかなラベル付けがされている場合でも、明瞭な地面の詳細（植生や道路など）を精緻化することができる。一般に、Paraformerは、利用可能なLR履歴ラベルから大規模なHR土地被覆マップを頑健に更新する可能性を示している。

4.4. アブレーション実験

この節では、Paraformerのさまざまな機能を評価するために、チェサピーク湾のデータセットに対してアブレーション実験を行った。表3の各アブレーションは以下のように説明される：(1)単独のCNNブランチは、LRラベルでCE損失を計算することにより依存的に学習される。(2)単独のTrans-formerブランチは、CNNブランチからの特徴の代わりにHR画像を埋め込み、LRラベルでCE損失を計算する。(3)PLATを含まないハイブリッド構造は、LRラベルで直接CE損失を計算する。

PLATモジュールを除去した結果、平均mIoUは62.81%となり、Paraformerの64.65%と比較して1.84%減少した。CNNブランチとTransformerブランチを除去することで、CNNブランチのみの結果は60.15%のmIoUとなり、4.5%減少した。唯一のTransformer分岐の結果は、56.49%の最も低いmIoUとなり、最も明らかな減少（8.16%）を示した。図8はParaformerの異なる出力を示している、

ここで、不正確なLRラベルはフレームワーク学習中に徐々に改良される。図8(g)に示す最終的な結果は、地上の詳細と、地上の真実と一致する正確な土地被覆パターンの両方を示している。さらに、図9は、CNNブランチ、Transformerブランチ、およびハイブリッド構造によってキャプチャされたコンテキストを視覚化したものである。図9(b)は、CNNブランチが、主に局所的な詳細（例えば、道路の端、一軒の家、低木）を捉えることに重点を置いていることを示している。図9(c)は、Trans-formerブランチが、ビル街や駐車場などの無傷の土地オブジェクトに焦点を当て、オブジェクトスケールの特徴を捉えていることを示している。ハイブリッド構造は、細かいエッジと無傷の領域の両方を持つ明白なオブジェクトに強い反応を示す。

一般的に、アブレーションの結果は2つの所見を示している：

(1) PLATモジュールは、大規模なHR土地被覆マッピングプロセスにおいて、フレームワークの学習を安定的に最適化し、LRラベルを合理的に利用することができる。**(2)**パラレルCNNとTransformerブランチはフレームワークの不可欠な部分であり、ローカルとグローバルのコンテキスト情報を橋渡しする、よりロバストな特徴抽出を構築する。

5. 結論

本論文では、大規模なHR土地被覆マップをHRラベルフリーでエンドツーエンドで更新するために、弱教師付きCNN-Transformerフレームワークである Paraformerを提案する。2つのデータセットを用いた実験により、Paraformerは、LR土地被覆データへのアクセスが容易な大規模なHRリモートセンシング画像のセマンティックセグメンテーションを導く上で、他のアプローチを凌駕する性能を示す。さらに分析を進めると、Paraformerは広域の様々な地形に頑健に適応し、正確なHR土地被覆マップを作成するために異なるLRラベルを安定的に利用できることが明らかになった。アブレーション研究により、並列CNN-Transformer構造とPLATモジュールの有効性が実証された。さらに、Paraformerの構成要素を透過的に説明するために、各トレーニングプロセスの中間結果と各ブランチの可視化されたコンテキストを示す。一般に、提案されたParaformerは、大規模なHR土地被覆マッピングを容易にする効果的な手法となる可能性を秘めている。

謝辞

本研究は、中国国家重点研究開発プログラム（助成金番号2022YFB3903605）および中国国家自然科学基金（助成金番号42071322）の支援を受けている。